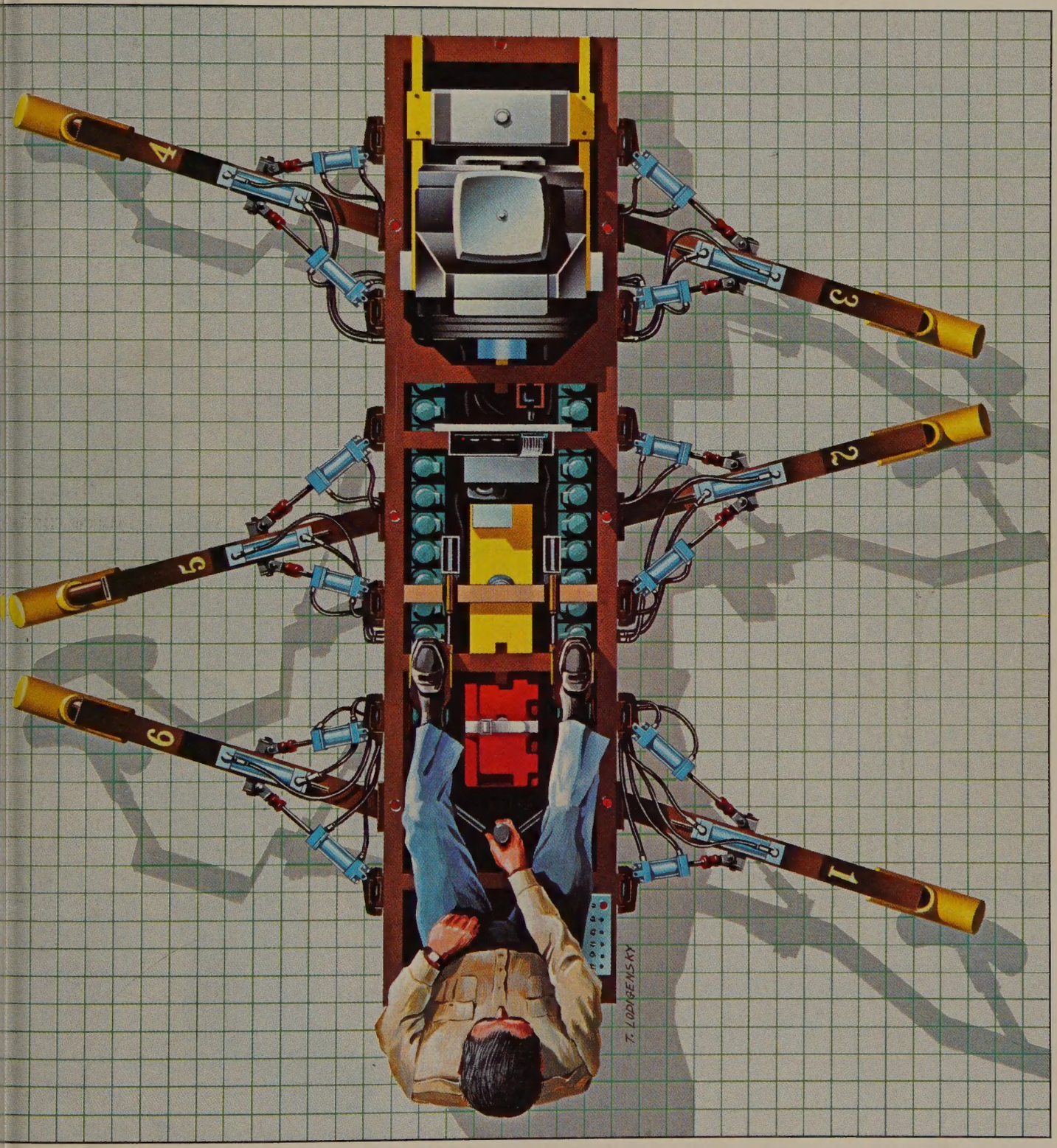
NOV 30 '83

S3
V.248
1983
(Jan.-Jun

SCIENTIFIC AMERICAN



MACHINES THAT WALK

$2.50

January 1983

74820 08715 01

Information, please.

Or, how the IBM Personal Computer can bring you the world.

Modem shown not supplied by IBM.

Established 1845 **SCIENTIFIC AMERICAN** January 1983 Volume 248 Number 1

ARTICLES

36 **THE LARGE-SCALE CULTIVATION OF MAMMALIAN CELLS, by Joseph Feder and William R. Tolbert** It is needed for the industrial production of clinically important proteins.

44 **MACHINES THAT WALK, by Marc H. Raibert and Ivan E. Sutherland**
Control technology makes it possible to build machines that can move like animals with legs.

54 **THE HAIR CELLS OF THE INNER EAR, by A. J. Hudspeth**
These biological transducers convert tiny mechanical forces into electrical signals to the brain.

74 **THE FOOTPRINTS OF EXTINCT ANIMALS, by David J. Mossman and William A. S. Sarjeant** Most species of extinct land vertebrates are known not from bones but from tracks.

86 **NMR SPECTROSCOPY OF LIVING CELLS, by R. G. Shulman**
Nuclear magnetic resonance (NMR) makes it possible to follow metabolism while it is in progress.

94 **THE PHYSICS OF ORGAN PIPES, by Neville H. Fletcher and Suszanne Thwaites**
The column of air resonating inside the pipe interacts with a jet blowing across the pipe's mouth.

104 **THE ORIGIN OF THE COSMIC X-RAY BACKGROUND, by Bruce Margon**
The diffuse glow of X-radiation that fills the universe may come from a multitude of quasars.

120 **THE MASS PRODUCTION OF IRON CASTINGS IN ANCIENT CHINA, by Hua Jue-ming** By 500 B.C. it was done by stack casting, in which iron is poured into multiple molds.

DEPARTMENTS

7 LETTERS
8 50 AND 100 YEARS AGO
10 THE AUTHORS
14 METAMAGICAL THEMAS
27 BOOKS
66 SCIENCE AND THE CITIZEN
130 THE AMATEUR SCIENTIST
138 BIBLIOGRAPHY

BOARD OF EDITORS Gerard Piel (Publisher), Dennis Flanagan (Editor), Brian P. Hayes (Associate Editor), Philip Morrison (Book Editor), Francis Bello, John M. Benditt, Peter G. Brown, Michael Feirtag, Jonathan B. Piel, John Purcell, James T. Rogers, Armand Schwab, Jr., Joseph Wisnovsky

ART DEPARTMENT Samuel L. Howard (Art Director), Steven R. Black (Assistant Art Director), Ilil Arbel, Edward Bell

PRODUCTION DEPARTMENT Richard Sasso (Production Manager), Carol Hansen and Leo J. Petruzzi (Assistants to the Production Manager), Carol Eisler (Senior Production Associate), Karen O'Connor (Assistant Production Manager), Carol Albert, Lori Mogol, Martin O. K. Paul, Julio E. Xavier

COPY DEPARTMENT Sally Porter Jenks (Copy Chief), Debra Q. Bennett, Mary Knight, Dorothy R. Patterson

GENERAL MANAGER George S. Conn

ADVERTISING DIRECTOR C. John Kirby

CIRCULATION MANAGER William H. Yokel

SECRETARY Arlene Wright

SCIENTIFIC AMERICAN (ISSN 0036-8733), PUBLISHED MONTHLY BY SCIENTIFIC AMERICAN, INC., 415 MADISON AVENUE, NEW YORK, N.Y. 10017. COPYRIGHT © 1982 BY SCIENTIFIC AMERICAN, INC. ALL RIGHTS RESERVED. PRINTED IN THE U.S.A. NO PART OF THIS ISSUE MAY BE REPRODUCED BY ANY MECHANICAL, PHOTOGRAPHIC OR ELECTRONIC PROCESS, OR IN THE FORM OF A PHONOGRAPHIC RECORDING, NOR MAY IT BE STORED IN A RETRIEVAL SYSTEM, TRANSMITTED OR OTHERWISE COPIED FOR PUBLIC OR PRIVATE USE WITHOUT WRITTEN PERMISSION OF THE PUBLISHER. SECOND-CLASS POSTAGE PAID AT NEW YORK, N.Y., AND AT ADDITIONAL MAILING OFFICES. AUTHORIZED AS SECOND-CLASS MAIL BY THE POST OFFICE DEPARTMENT, OTTAWA, CANADA, AND FOR PAYMENT OF POSTAGE IN CASH. SUBSCRIPTION RATE: $21 PER YEAR IN THE U.S. AND ITS POSSESSIONS; $27 PER YEAR IN ALL OTHER COUNTRIES. POSTMASTER: SEND ADDRESS CHANGES TO SCIENTIFIC AMERICAN, 415 MADISON AVENUE, NEW YORK, N.Y. 10017.

SCIENTIFIC AMERICAN

CORRESPONDENCE

Offprints of more than 1,000 selected articles from earlier issues of this magazine, listed in an annual catalogue, are available at 75 cents each. Correspondence, orders and requests for the catalogue should be addressed to W. H. Freeman and Company, 660 Market Street, San Francisco, Calif. 94104. Offprints adopted for classroom use may be ordered direct or through a college bookstore. Student sets of 10 or more Offprints are collated by the publisher and are delivered as sets to bookstores.

Photocopying rights are hereby granted by Scientific American, Inc., to libraries and others registered with the Copyright Clearance Center (CCC) to photocopy articles in this issue of SCIENTIFIC AMERICAN for the flat fee of $2 per copy of each article or any part thereof. Such clearance does not extend to the photocopying of articles for promotion or other commercial purposes. Correspondence and payment should be addressed to Copyright Clearance Center, Inc., 21 Congress Street, Salem, Mass. 01970. Specify CCC Reference Number ISSN 0036–8733/83. $2.00 + 0.00.

Editorial correspondence should be addressed to The Editors, SCIENTIFIC AMERICAN, 415 Madison Avenue, New York, N.Y. 10017. Manuscripts are submitted at the authors' risk and will not be returned unless they are accompanied by postage.

Advertising correspondence should be addressed to C. John Kirby, Advertising Director, SCIENTIFIC AMERICAN, 415 Madison Avenue, New York, N.Y. 10017.

Subscription correspondence should be addressed to Subscription Manager, SCIENTIFIC AMERICAN, P.O. Box 5969, New York, N.Y. 10017. The date of the last issue on your subscription is shown in the upper right-hand corner of each month's mailing label. For change of address notify us at least four weeks in advance. Please send your old address (if convenient, on a mailing label of a recent issue) as well as the new one.

Name

New Address

Street

City

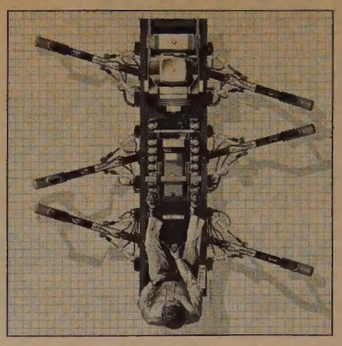
State and ZIP

Old Address

Street

City

State and ZIP



THE COVER

The painting on the cover depicts a six-legged machine that walks, largely under the control of a built-in microprocessor (see "Machines That Walk," by Marc H. Raibert and Ivan E. Sutherland, page 44). More precisely, the machine crawls, meaning that it employs a method of locomotion that does not demand attention to the problem of balance. At least three of the six legs always touch the ground to provide a tripod for support. The microprocessor operates on data supplied by sensing devices that report to it the position of each leg and the forces on the leg. The human driver can steer the machine by making the legs on one side move faster than those on the other. He also can adjust the attitude of the body and its height above the ground and can aid the microprocessor in choosing suitable footholds on rough terrain. Power is provided by an 18-horsepower gasoline engine; hydraulic actuators move the legs. The driver sits in back where he can see the legs. The potential of machines that walk is that they can function in terrain where vehicles with wheels or tracks are ineffective.

THE ILLUSTRATIONS

Cover painting by Ted Lodigensky

Page	Source	Page	Source
22	Lee Sallows	98–99	Andrew Christie
37	Jack C. Wahl, Monsanto Company	100	Pieter Stroethoff
		101	Neville H. Fletcher
38–42	Jerome Kuhl	102–103	Andrew Christie
43	Jerome Kuhl (*top*); Joseph Feder and William R. Tolbert, Monsanto Company (*bottom*)	105	© 1960 National Geographic Society–Palomar Observatory Sky Survey. Reproduced by permission of the California Institute of Technology (*top*); Gary A. Chanan, Columbia University (*bottom*)
45	Leg Laboratory, Carnegie-Mellon University		
46–51	George V. Kelvin		
52	R. F. Bonifield		
53	R. F. Bonifield, courtesy of New York Public Library	108	Ilil Arbel
55	Richard A. Jacobs, California Institute of Technology	109	Walken Graphics
		110	© 1960 National Geographic Society–Palomar Observatory Sky Survey. Reproduced by permission of the California Institute of Technology (*top*); Gary A. Chanan, Columbia University (*bottom*)
56	Bunji Tagawa		
57	A. J. Hudspeth		
58–60	Bunji Tagawa		
61	A. J. Hudspeth		
62–64	Bunji Tagawa		
74	Suzanne Swibald		
76–81	Patricia J. Wynne	113	Bruce Margon, University of Washington (*left*); Ilil Arbel (*right*)
82	Philip J. Currie (*top*), Giuseppe Leonardi (*bottom*)		
84	American Museum of Natural History	114–115	Ilil Arbel
		120–122	Michael Goodman
87–93	Gabor Kiss	123	Hua Jue-ming
95	Pieter Stroethoff	124–126	Michael Goodman
96	Andrew Christie	127	Hua Jue-ming
97	Neville H. Fletcher	128–136	Michael Goodman

WHAT'S BETTER THAN SPEED READING?

SPEED LEARNING

(SPEED PLUS COMPREHENSION)

Speed Learning is replacing speed reading. It's easy to learn...lasts a lifetime...applies to everything you read...and is the only accredited course with the option of college or continuing education credits.

Do you have too much to read and too little time to read it? Do you mentally pronounce each word as you read? Do you frequently have to go back and re-read words or whole paragraphs you just finished reading? Do you have trouble concentrating? Do you quickly forget most of what you read?

If you answer "yes" to any of these questions — then here at last is the practical help you've been waiting for. Whether you read for business or pleasure, school or college, you will build exceptional skills from this major breakthrough in effective reading, created by Dr. Russell Stauffer at the University of Delaware.

Not just "speed reading" — but speed reading-thinking-understanding-remembering-and-learning

The new *Speed Learning Program* shows you step-by-proven-step how to increase your reading skill and speed, so you understand more, remember more and use more of everything you read. The typical remark made by the 75,000 slow readers who completed the *Speed Learning Program* was: "Why didn't someone teach me this a long time ago?" They were no longer held back by the lack of skills and poor reading habits. They could read almost as fast as they could think.

What makes Speed Learning so successful?

The new *Speed Learning Program* does not offer you a rehash of the usual eye-exercises, timing devices, costly gadgets you've probably heard about in connection with speed reading courses or even tried and found ineffective.

In just a few spare minutes a day of easy reading and exciting listening, you discover an entirely new way to read and think — a radical departure from anything you have ever seen or heard about. Research shows that reading is 95% *thinking* and only 5% eye movement. Yet most of today's speed reading programs spend their time teaching you rapid eye movement (5% of the problem) and ignore the most important part (95%) *thinking*. In brief, *Speed Learning* gives you what speed reading *can't*.

Imagine the new freedom you'll have when you learn how to dash through all types of reading material *at least* twice as fast as you do now, and with greater comprehension. Think of being able to get on top of the avalanche of newspapers, magazines and correspondence you have to read . . . finishing a stimulating book and retaining facts and details more clearly and with greater accuracy than ever before.

Listen-and-learn at your own pace

This is a practical, easy-to-learn program that will work for you — no matter how slow a reader you think you are now. The *Speed Learning Program* is scientifically planned to get you started quickly . . . to help you in spare minutes a day. It brings you a "teacher-on-cassettes" who guides you, instructs, encourages you, explaining material as you read. Interesting items taken from *Time* Magazine, *Business Week*, *Wall Street Journal*, *Family Circle*, *N.Y. Times* and many others, make the program stimulating, easy and fun . . . and so much more effective.

Executives, students, professional people, men and women in all walks of life from 15 to 70 have benefited from this program. Speed Learning is a fully accredited course . . . costing only 1/5 the price of less effective speed reading classroom courses. Now you can examine the same, easy, practical and proven methods at home . . . in spare time . . . without risking a penny.

Examine Speed Learning FREE for 15 days

You will be thrilled at how quickly this program will begin to develop new thinking and reading skills. After listening to just one cassette and reading the preface you will quickly see how you can achieve increases in both the speed at which you read and in the amount you understand and remember.

You must be delighted with what you see or you pay nothing. Examine this remarkable program for 15 days. If, at the end of that time you are not convinced that you would like to master Speed Learning, simply return the program and owe nothing. See the coupon for low price and convenient credit terms.

Note: Many companies and government agencies have tuition assistance plans for employees providing full or partial payment for college credit programs.

In most cases, the entire cost of your Speed Learning Program is Tax Deductible.

COLLEGE CREDITS

You may obtain *2 full semester hour credits* for course completion, wherever you reside. Credits offered through Whittier College (California). Details included in your program.

CONTINUING EDUCATION UNITS

National Management Association, the world's largest association of professional managers, awards 3.0 CEU's for course completion. CEU's can be applied toward the certificate in Management Studies.

PROFESSIONAL SOCIETIES

Speed Learning is offered internationally to members of professional associations such as: American Chemical Society, Foundation for Accounting Education, Institute of Electrical and Electronics Engineers and dozens more. Consult your Education Director for information.

BUSINESS, INDUSTRY, GOVERNMENT

Many companies and government agencies offer Speed Learning as a wholly-paid or tuition reimbursement program. Consult your Training or Personnel Director for details.

DZ-21

learn INCORPORATED 113 Gaither Drive, Mt. Laurel, NJ 08054

YES! Please rush me the materials checked below:

- ☐ Please send the Speed Learning program @ $99.95 plus $4 postage and handling.
- ☐ Please send the Speed Learning Medical Edition @ $109.95 plus $4 postage and handling.
- ☐ Please send the Junior Speed Learning program (ages 11 to 16) @ $89.95 plus $4 postage and handling.

NJ residents add 5% sales tax.

Check method of payment below:

- ☐ Check or money order enclosed (payable to learn incorporated)
- ☐ Charge my credit card under the regular payment terms
  - ☐ Visa ☐ Master Card Interbank No._____ ☐ American Express
  - Card No._____ Exp. Date_____

I understand that if after 15 days I am not delighted in every way, that I may return the materials in their original condition for a full refund. No questions asked.

Name _____

Address _____

City_____ State_____ Zip_____

Signature_____

If you don't already own a cassette player, you may order this Deluxe Cassette Recorder for only $49.95. (Includes handling and delivery.)

Check here to order ☐

Outside USA add $10 per item—Airmail extra

SCIENCE/SCOPE

Two communications satellites made history as the first to be launched from NASA's space shuttle. The first, SBS-3, is operated by Satellite Business Systems and will carry high-speed data for many U.S. companies. The second, Anik-C, is operated by Telesat Canada and will improve telephone, television, and data service in Canada. The satellites are versions of Hughes Aircraft Company's HS 376, the world's most widely purchased communications satellite. Hughes now has built 70% of the world's operating commercial communications satellites and has more successes than all other companies combined.

A safety device that snuffs out explosions in the blink of an eye, originally developed for the military, is being applied in commercial situations where fire poses an immediate threat to human life. The Dual Spectrum™ sensing and suppression system has been evaluated in New York Transit Authority toll booths. It detects fire bomb explosions set off by criminals, and suppresses them in one-tenth of a second -- before transit employees can be injured. The system could be applied almost anywhere fire explosions occur within an enclosed area. It was developed by the Santa Barbara Research Center, a Hughes subsidiary.

A new "quick draw" capability for the Maverick missile system would let pilots hit more targets in less time and reduce their risk of being hit by enemy fire. In the last of 21 flight tests, a U.S. Air Force F-4 fighter crew fired three air-to-surface Mavericks within 12 seconds from an altitude of 700 feet. The three missiles scored direct hits on three trucks parked about 70 meters apart in normal convoy fashion. The tests employed a three-rail launcher with a modified electrical system. Although TV-guided versions of the Hughes missile were used in the tests, the system could be used for imaging infrared Mavericks.

Advanced military electro-optical systems are being produced in large numbers and at high rates at a new Hughes manufacturing facility. The complex, which covers one-half million square feet, is designed specifically for making such high-technology devices as infrared night sights and laser rangefinders. Recent milestones include the following deliveries to the U.S. Army: the 2,000th laser tank fire control system for the M60A3 tank, the 1,000th airborne TOW antitank missile system for the Cobra attack helicopter, and the 1,000th thermal imaging system and laser rangefinder for the M1 Abrams tank. In addition, production rates for the two M1 units have reached 70 per month.

Hughes Radar Systems Group has career opportunities for engineers, scientists, and programmers. We design and build many of today's most complex airborne and spaceborne radar electronics systems, including data links, electronic warfare systems, and display systems. We need systems analysts, microwave specialists (antenna, receivers, transmitters, data processors), circuit designers (analog, digital, RF/IF), scientific programmers, mechanical designers, systems and test engineers. Send resume to Engineering Employment, Dept. SSA, Hughes Radar Systems Group, P.O. Box 92426, Los Angeles, CA 90009. Equal opportunity employer.

Creating a new world with electronics

HUGHES

HUGHES AIRCRAFT COMPANY
CULVER CITY, CALIFORNIA 90230

(213) 670-1515 EXTENSION 5964

LETTERS

Sirs:

In David L. Waltz's article "Artificial Intelligence" [SCIENTIFIC AMERICAN, October, 1982] he attributes the name of Terry Winograd's computer language—SHRDLU—to the seventh through the 12th most frequent letters in the English language. Of course they are not. The generally accepted progression is E, A, O, I, D, H, N, R, S, T, U, Y....

The origin of the character string SHRDLU is interesting. When Linotype operators made a mistake, they marked the slug of type by running their fingers down the leftmost and second leftmost keys of the machine's keyboard, producing the characters ETAOIN SHRDLU. This was a signal to the compositor that the line was faulty and was to be discarded. Occasionally the bad slug slipped past the compositor and the proofreader and inexplicably appeared in the middle of a newspaper story. It was used as the name of a character in several humorous writings, one by James Thurber, who was a newspaperman before he became an essayist and a cartoonist.

THOMAS BAKERSMITH

Maplewood, Mo.

Sirs:

As a physicist and an avid timpanist I read Thomas D. Rossing's article "The Physics of Kettledrums" [SCIENTIFIC AMERICAN, November, 1982] with great interest. His description of air-mass loading can be used to explain one phenomenon familiar to timpanists: the decrease in pitch for loud kettledrum notes. The louder the note, the greater the effects of air-mass loading and thus the lower the pitch. Hence when a timpanist is tuning, he must take into consideration the loudness of the note to be played. For instance, in the scherzo movement of Beethoven's Ninth Symphony there is a prominent three-note timpani phrase, played first very loud, then very soft. Between these phrases careful timpanists change the tension in the drumhead to compensate for air-mass loading, thereby ensuring that the phrases sound at the same pitch.

Whereas the pitch of most instruments rises with the loudness of the note, the kettledrum is one of the very few for which louder notes become lower in pitch, owing to air-mass loading.

DAVID G. STORK

Swarthmore College
Swarthmore, Pa.

Sirs:

In connection with F. Richard Stephenson's article "Historical Eclipses" [SCIENTIFIC AMERICAN, October, 1982] it may be of interest that a lunar eclipse and a solar eclipse both occurring in one month at the end of the year 1044 were the subject of a poem by Samuel HaNagid (the Prince), a Talmudist, poet and statesman, vizier of Granada during the peculiar golden period of peace between Muslims and Jews in 11th-century Spain. This period and the life of Samuel and of Ibn Gabirol, the poet and philosopher, are fictionalized in *The Green Rose,* by Warren A. Silver (Dial Press, 1977). HaNagid was recognized by his Arab contemporaries as a mathematician and astronomer. Some lines of his poem read in part (translation by T. Carmi, *The Penguin Book of Hebrew Verse,* Viking Press, 1981):

"He who is master of might and beauty, He struck both his luminaries in the very same month. He covered the face of the moon with His terrestrial globe and blocked off the sun with His moon."

MALCOLM S. MASON

Washington, D.C.

Sirs:

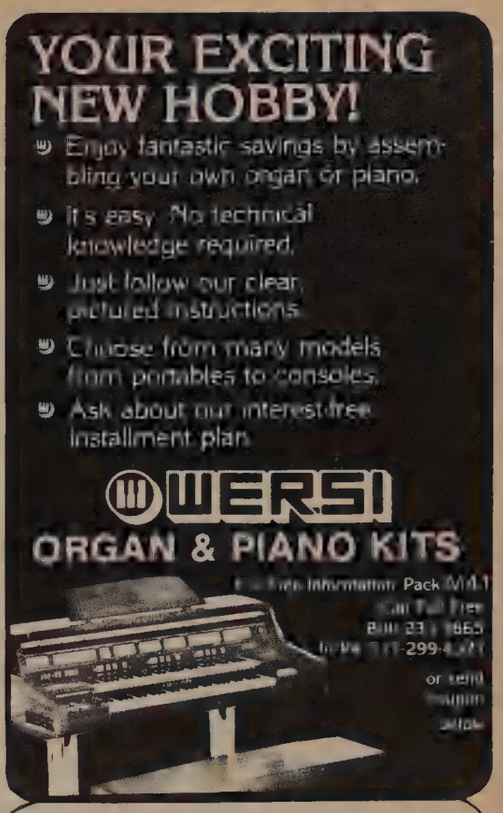
The August 11, 1124, solar eclipse, reported as total in Russia, is reported as partial in Flanders as follows:

"In the year 1124 from the Incarnation of our Lord, in the month of August, there was visible to all the inhabitants of the lands an eclipse on the body of the sun at about the ninth hour of the day, and an unnatural failure of light so that the eastern part of the sun, darkened little by little, poured forth strange clouds on the other parts, not darkening the whole sun at the same time, however, but only partially. Nevertheless, the same cloud wandered over the whole circle of the sun, moving across from east to west, but only within the circle of the solar essence."

The reporter is Galbert, a notary in Bruges, writing in the year 1127 or 1128 in his book *The Murder of Charles the Good, Count of Flanders* (Harper & Row, 1967). Because Bruges and Kiev, from where I presume the Russian report originates, are within one degree of latitude from each other, either the difference must stem from the 27 degrees separating them in longitude or the Russian report might be an exaggeration.

LEONARDAS V. GERULAITIS

Associate Professor
Department of History
Oakland University
Rochester, Mich.



WERSI Dept. M41 P.O. Box 5318
Lancaster, PA 17601

☐ Free Info. Pack: ☐ Organ ☐ Piano
☐ Catalog & Demo Record *(enclose $3)*

Name _____
Address _____
City _____ State _____ Zip _____
Phone ( ) _____

Rep inquiries invited

Free Guidebook To Winter Wonders.

Our free 30-page book offers a warm, inviting look at some very special vacation opportunities at Kiawah Island. Come play golf or tennis, savor superb Lowcountry cuisine, enjoy shopping, jeep safaris, our ten miles of breathtaking beach, and all the adventure of nearby historic Charleston. Off-season rates now make our island escapes a better value than ever. For reservations, call us or your travel agent.

Name _____
Address _____
City/State/Zip _____

VMSI 1

Kiawah Island
P.O. Box 12910
Charleston, SC 29412
CALL TOLL FREE 1-800-845-2471.
(In S.C., 803-768-2121.)

50 AND 100 YEARS AGO

SCIENTIFIC AMERICAN

JANUARY, 1933: "The most recent atomic research offers the heartening possibility of ultimately amalgamating the wave and particle concepts of matter. As a matter of fact, there are *no* experiments which prove that matter possesses *all* the properties of *either* a wave *or* a particle. Accordingly Werner Heisenberg has suggested that the two mental pictures that experiment conjures out of our imagination—one of particles, the other of waves—are both to be viewed simply as incomplete analogies arising from our temporary inability to describe matter in everyday language. Although we cannot draw a satisfactory picture of the atom as it is conceived by wave mechanics, the mathematics of the theory enables us to do all the things that were possible with the Bohr model of the atom. In addition it appears that what were heretofore contradictions are removed, fundamental points are refined and made more precise, and the number of assumptions necessary to attain those ends are reduced to a minimum. From this point of view wave mechanics signifies not so much a radical change as it does a welcome and highly significant evolution of the existent atomic theory."

"Photographs at the rate of 4,000 a second with exposures ranging from 1/100,000 to 1/500,000 of a second have been made at the Massachusetts Institute of Technology by means of a unique electrical circuit that produces light with great actinic intensity. The instantaneous flash of this light is many times more brilliant than the sun. The new circuit, which employs either mercury-arc tubes or spark gaps, was developed by Professor Harold E. Edgerton and Kenneth J. Germeshausen of the Department of Electrical Engineering. The device has already been used to make striking photographs in which familiar things are shown in astonishing new forms. The splash of a drop of milk falling into milk is revealed in the shape of a miniature crown tipped with infinitesimal pearl-like drops. The photographic image recorded at the instant of impact between a golf club and a ball shows clearly the momentary flattening of the latter at the point of contact."

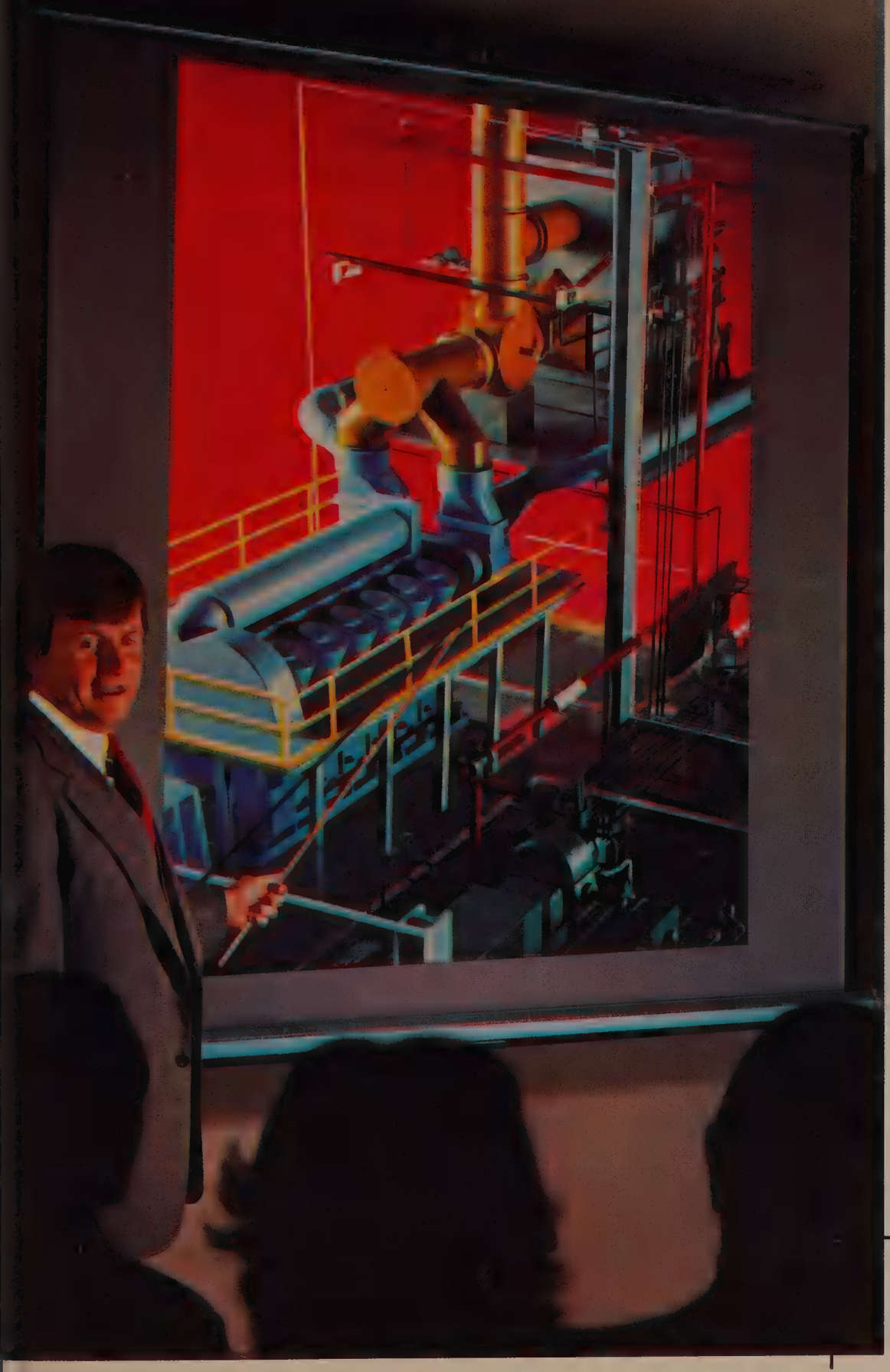
"The plant patent law passed a little more than two years ago has resulted to date in the granting of 39 patents on flowers, fruits and other plant life. Luther Burbank, the famous plant breeder, although dead, has received six patents through his executrix, Elizabeth Burbank of Santa Rosa, Calif. Burbank now has to his credit more plant patents than any other plant breeder. Last April four patents were issued to him covering a new red freestone plum, a pure golden freestone plum, a large, beautifully colored variety of deep carmine plum and a yellow freestone peach. In May he received another patent for a red-skinned, golden-fleshed plum, very large in size. His sixth patent, issued in July, was for a chrome yellow, apricot-like flavored plum."

"The qualifications for the new profession of air hostess are quite severe. The young lady has to be single, under 25, a college graduate and weigh not more than 125 pounds. Apparently the hostess has to move about, and must be light enough not to unbalance the ship! A three-hour flight test for air-sickness and two weeks' observation by pilots are other hurdles. The duties are to attend to passengers, entertain them and divert their minds from the imaginary dangers of a first flight. It is significant of the enterprise and business ability of our transport operators that they do provide such service."

JANUARY, 1883: "A petroleum motor, or rather an engine for obtaining motor power from an explosive mixture of gasoline vapor and air, has been constructed by a firm in Hanover. The working cylinder is eight inches in diameter, with a $14\frac{3}{16}$-inch stroke. The design of the machine is similar to that of a type of gas engine constructed by Wittig and Hees. The gasoline is led through a pipe to the pump cylinder, where it mixes with a definite proportion of atmospheric air. The mixture is then compressed and forced into the working cylinder, where it is ignited by a lamp separately supplied with oil. In four trials with the particular engine in question the maximum force obtained was 4.5 horse power, with 130 revolutions per minute. The consumption of spirit of specific gravity 0.675 was at the rate of $1\frac{3}{4}$ to $2\frac{1}{2}$ pints per horse power per hour. The machine is stated to require as little attention and to work as cheaply as a gas engine."

"The use of dynamo machines in the generation of electricity for public information has now been sufficiently demonstrated to be both practical and economical, especially in large areas where considerable light is required, such as railway stations, harbors and public institutions, not to mention lighthouses, in which the electric light has been used to advantage for a number of years. Next to its importance in the production of the electric light, probably the most useful purpose to which the dynamo has yet been applied is the transmission of power. This property depends on the principle known as the *reversibility* of the dynamo, or, in other words, the transformation of electrical energy into mechanical work, which is simply the converse of converting mechanical into electrical energy. In the transmission of power by electricity the current is generated in one machine, from which it is conveyed to another, which may be at any distance, and may be utilized to set machinery of any kind in motion. The distance over which the electric transmission of power may be carried without serious loss of energy is still a matter of controversy, not having yet been proved by works of any great magnitude. We have it, however, on the authority of Dr. William Siemens, as the result of a long series of experiments made by this distinguished electrician, that the total loss resulting from the double process of conversion (*i.e.*, the conversion of mechanical into electrical energy and secondly the conversion of electrical into mechanical work) is about 20 per cent. To this loss should be added that caused by the resistance of the wires, which depends on their length and sectional area; so that for actual work it is considered safe to assume that the loss sustained over a considerable distance does not exceed 50 per cent., which will not appear very great when compared with the loss sustained in transmitting power by compressed air, water or other means."

"The apparatus employed for separating the various ingredients contained in crude petroleum in the process of obtaining kerosene consists of an iron still having a wrought-iron worm pipe, which is submerged in a tank containing cold water. The still having been filled with crude oil, a fire lighted under it causes the oil to boil and drives off the more volatile vapors. By surrounding the coil with ice, or by compressing these gases by means of an air pump, they may be condensed into the form of very volatile liquids. The average yield of crude Pennsylvania oil is stated to be gasoline $1\frac{1}{2}$, refined naphtha 10, benzine 4, refined petroleum or kerosene 55, lubricating oil $17\frac{1}{2}$, paraffine 2, loss gas and coke 10, total 100. By slow distillation in high stills the production of the heavier oils can be avoided, they being 'cracked' into lighter oils; only crude naphtha, kerosene and coke result."



Introducing Polaroid's Instant Overhead Transparency.

Colorful graphs. Multihued maps. Brilliant photographs.

You can now make management presentations with the powerful impact of color visuals like these, simply, inexpensively, and *instantly.*

Polaroid's Colorgraph 8 × 10 Instant Overhead Transparency System provides you with transparencies that can make your presentations look as professional as you are. And when your information is presented in such a colorful, graphic format, it's bound to be more clearly understood and better remembered. Because color is one of the most effective tools in visual communication.

Besides which, your transparencies can be developed in just 4 minutes. Which means you can present the latest information at the earliest possible moment. Or you can make important changes *moments* before a presentation is scheduled.

What's more, you don't have to send out for your Colorgraph transparencies. You can send in. To the security of your own in-house photo or AV departments. Where your transparencies can be produced easily and quickly from 35mm slides or original art.

The Colorgraph illustrated to the left was made from a 35mm slide in just 4 minutes using Polaroid's new 8 × 10 Polaprinter Instant Slide Copier.

Now full color business presentations can be ready in 4 minutes.

For more information and a demonstration return this coupon to: Polaroid Corp., Industrial Marketing Dept. 553, P.O. Box 5011, Clifton, N.J. 07015. Or call toll free in continental U.S.: (800) 526-7843, Ext. 400. In N.J. call: 800-522-4503, Ext. 400.

Name _____

Title _____

Company _____

Address _____

City _____

State _____ Zip _____

Telephone _____

POLAROID

Plant model courtesy of Thermo Electron Corp. Waltham, MA © 1982 Polaroid Corp. "Polaroid," "Colorgraph" and "Polaprinter"®

NordicTrack

Jarless Total Body Cardiovascular Exerciser Duplicates X-C Skiing for the Best Motion in Fitness

The enjoyable sport of cross-country skiing is often cited by physiologists as the most perfect form of cardiovascular exercise for both men and women. Its smooth, fluid, total body motion uniformly exercises more muscles and higher heart rates seem easier to attain than when jogging or cycling. NordicTrack closely simulates the pleasant X-C skiing motion and provides the same cardiovascular endurance-building benefits—right in the convenience of your home. Makes a year round, consistent exercise program easily attainable. Eliminates the usual barriers of time, weather, chance of injury, etc. Also highly effective for weight control.

Better Than Running
NordicTrack gives you a more complete work out—conditions both upper body and lower body muscles at the same time. Fluid, jarless motion will not cause joint or back pains as jogging or running often does.

More Effective Than Exercise Bikes
NordicTrack's stand-up skiing motion more uniformly exercises large leg muscles and also adds important upper body exercise. Higher pulse rates, necessary for building fitness, seem easier to attain because the work is shared by more muscle mass. The NordicTrack also keeps more muscles in tone.

Even Better Than Swimming
NordicTrack more effectively exercises the largest muscles in the body, those located in the legs and buttocks. These muscles' main function is to lift and propel the body in the standing position. However, when swimming, the body is supported by the water, thus preventing these muscles from being effectively exercised. The stand up exercising position on the NordicTrack much more effectively exercises these muscles.

A Proven, High Quality Durable Product
Ruggedly constructed, NordicTrack is quiet, motorless and folds compactly for storage. Separately adjustable arm and leg resistances. NordicTrack is in its 6th year of production and is used in thousands of homes and many institutions. We manufacture and sell direct. One year warrantee, 15 day trial period with return privilege.

Call or Write for FREE BROCHURE
TOLL FREE 1-800-328-5888
8 AM-5 PM MON.-FRI.
PSI 124 F Columbia ct. Chaska, MN 55318
Minnesota 612-448-5987

THE AUTHORS

JOSEPH FEDER and WILLIAM R. TOLBERT ("The Large-Scale Cultivation of Mammalian Cells") are members of the research department of the Monsanto Company who have worked together on the subject of their article for the past eight years. Feder is senior science fellow at Monsanto. He was graduated in 1953 with a B.S. from Roosevelt University. He went on to obtain his M.S. (1961) and his Ph.D. in biochemistry (1964) from the Illinois Institute of Technology. He is adjunct professor of biochemistry at the University of Missouri at St. Louis. Tolbert is research group leader at Monsanto. His B.S. (1964) is from the University of Richmond. His M.S. in physics (1966) and Ph.D. in biophysics (1971) are from the University of Wisconsin at Madison. In 1971 and 1972 he was research associate at Duke University. From 1972 to 1974 he was postdoctoral fellow at the Allegheny General Hospital; from there he moved to Monsanto. Tolbert describes himself as "an inveterate gadgeteer in [his] approach to scientific problems" whose interests outside the laboratory include the making of wine from fruit that he grows.

MARC H. RAIBERT and IVAN E. SUTHERLAND ("Machines That Walk") are respectively assistant professor of computer science and robotics and visiting scientist at Carnegie-Mellon University. They have collaborated and competed in the building of machines capable of walking since 1979. Raibert writes: "After finishing my Ph.D. at the Massachusetts Institute of Technology I went to the Jet Propulsion Laboratory of the California Institute of Technology, where I worked on a number of projects in robotics. It was during my stay at J.P.L. that I met Ivan Sutherland, who was then chairman of computer science at Cal Tech. Ivan and I began to talk about projects. I told him I thought we could learn a good deal about locomotion from pogo sticks. I was more than a little surprised when he showed genuine enthusiasm about the idea and coughed up $3,000 for me to build a first model. He himself began to spend substantial time in the shop building a small pneumatically powered walker." Sutherland, who has a Ph.D. from M.I.T. and has been affiliated with several universities and companies, adds: "I worked on robots during my college years but then got involved in other research until Marc Raibert revived my interest in locomotion in 1979. Since then I have been working about half time in robotics. The other half of my time I divide between work in the venture-capital business and the consulting work of Sutherland, Sproull and Associates."

A. J. HUDSPETH ("The Hair Cells of the Inner Ear") is professor of biology at the California Institute of Technology; later this year he will move to the University of California School of Medicine in San Francisco as professor of physiology. He has four degrees from Harvard: a B.A. from Harvard College (1967), an M.A. and a Ph.D. from Harvard University (1968 and 1973) and an M.D. from the Harvard Medical School (1974). In 1974 he was visiting research fellow at the Karolinska Hospital in Stockholm. In 1975 he moved to Cal Tech.

DAVID J. MOSSMAN and WILLIAM A. S. SARJEANT ("The Footprints of Extinct Animals") are geologists with a common interest in the subject of their article. Mossman is a native of Canada whose B.Sc. and M.Sc. are from Dalhousie University in Nova Scotia. He went to New Zealand to continue his education, getting his Ph.D. from the University of Otago. He returned to Canada to join the faculty at the University of Saskatchewan. He taught at Saskatchewan from 1971 to 1982, leaving to take up his current job as associate professor of economic geology at Mount Allison University in New Brunswick. Mossman's major interest is mineral deposits. Sarjeant was born and educated in England; his B.Sc. and Ph.D. are from the University of Sheffield. After serving briefly on the faculties of the University College of North Staffordshire and the University of Reading he went to the University of Nottingham in 1963 and remained there until 1972. In 1972 he emigrated to Canada to become professor of geological science at the University of Saskatchewan. Among his interests is palynology: the study of fossil spores and pollens.

R. G. SHULMAN ("NMR Spectroscopy of Living Cells") is professor of molecular biophysics and biochemistry and chemistry and director of the division of biological sciences at Yale University. He got his education at Columbia University, from which he has three degrees: an A.B. (1943), an A.M. (1947) and a Ph.D. in chemistry (1949). In 1949 and 1950 he was an Atomic Energy Commission fellow at the California Institute of Technology. From 1950 to 1953 he was head of semiconductor research at the Hughes Aircraft Company. From 1953 until 1979 he was at Bell Laboratories; at the end of that time he moved to Yale. In addition to the subject of his article, his research interests

You can see the Great Wall from the moon. Or see the moon from the Great Wall.

Fantastic though it seems the Great Wall of China is actually visible from the moon.

But Salén Lindblad Cruising offers you the chance to see it from a little closer perspective on our China Coastal/Yangtse River cruise aboard the **M.V. Yao Hua.***

You'll travel along the coast of China and up the Yangtse River spending three nights in a lovely Beijing hotel. You'll also visit Hong Kong, Shanghai, Xiamen, Nanjing, Suzhou—the places that make China one of the most intriguing spots under the sun. Or moon.

But this is also a culinary journey: along the way you will travel through China's four distinct cuisines as you travel through China itself. (Excellent Scandinavian food is also served.) And your entire journey will be accompanied by eminent Sinologists.

It's a living experience. A learning experience. One that will tear down walls —even as you visit the greatest wall of all.

Cruises aboard the **Yao Hua** leave weekly, March through November 1983. For more information send us the coupon or talk to your travel agent or call **800-223-5688**. In New York call **(212) 751-2300**.

*Registered in China.

Please send me your brochure on China Coastal/Yangtse River cruises aboard the **Yao Hua**.

Name

Address

City, State, Zip

Phone No.

Travel Agent SA-1

TO LEARN IS TO LIVE.

Salén Lindblad
Cruising, Inc.
133 East 55th Street
N.Y., N.Y. 10022 U.S.A.
Telex: 423661

SALÉN LINDBLAD CRUISING

Introducing

THE QUESTAR® M 1

the long-range microscope for non-destructive testing . . . non-disruptive observation . . . non-contact gauging and inspection

Questar optical systems have once again extended the range of the services they can perform, with the design of the M 1. This system was especially created to provide peak performance at distances between 22 and 77 inches.

At 22 inches the target is within arm's length of the technician should he have need to shift its position; as, for example, in observing live specimens in an aquarium. At the other extreme, hot or dangerous subjects may be viewed from a vantage point of 77 inches.

Magnifications range from 5 to 65 powers, depending on the distance and the eyepiece used. Five eyepieces, ranging from 32 mm. to 8 mm. are provided. At 22 inches with an 8 mm. eyepiece, the power is 65 diameters and this is almost doubled when, with the flick of a lever, you introduce an internally-mounted Barlow lens into the system.

With Questar's extreme resolution, excessive magnification is not necessary – the detail is all there to be seen. The M 1 on standard optical tests charts test to better than 130 lines per millimeter. However, in laboratory use at Los Alamos we are told that objects as small as three microns have been identified.

The Questar M 1 is photo-visual, of course, having provision for attaching a 35 mm. SLR camera or a high-speed video camera. Many applications exist here for monitoring and recording, from a remote position, operations which entail either delicate or hazardous conditions.

The original Questar system has had the ability to focus sharply from infinity to 10 feet, a unique and remarkable achievement. Now the M 1 provides a new way of looking at the micro world. It has already been used for such widely diverse jobs as detecting stress lines in propellers, inspecting minute connections on circuit boards, observing magnetically-oriented bacteria in their habitat, monitoring with a high-speed video camera the coating of a 200 micron nuclear target. In many applications engineers and scientists are finding the Questar M 1 the answer to problems they have long sought to solve. We would like to try to help you solve yours.

Literature with specifications on request.

©1982, Questar Corporation

Our booklet describing Questar, the world's finest, most versatile telescope, contains an astonishing collection of photographs by Questar owners. Send $2 for mailing costs on this continent; by air to S. America, $3.50; Europe and N. Africa, $4.00; elsewhere, $4.50.

QUESTAR

Box C, Dept. 203, New Hope, Pa. 18938
(215) 862-5277

include radiation damage to DNA, bacterial-virus genetics and paramagnetic metal-ion complexes of nucleic acids.

NEVILLE H. FLETCHER and SUSZANNE THWAITES ("The Physics of Organ Pipes") are respectively professor of physics and a graduate student in physics at the University of New England in Australia. Fletcher obtained two degrees from the University of Sydney: a B.Sc. and a D.Sc. Between getting them he earned his Ph.D. from Harvard University. His original research interests were solid-state physics and the physics of clouds, but in the past 10 years he has given much attention to musical acoustics and the mechanisms of sound production in animals. Fletcher is president of the Australian Institute of Physics and Secretary for Physical Sciences of the Australian Academy of Science. Thwaites has a B.Sc. and an M.Sc. from the University of Western Australia, where she worked on the formation of hail.

BRUCE MARGON ("The Origin of the Cosmic X-Ray Background") is professor of astronomy and chairman of the department of astronomy at the University of Washington. He received his undergraduate education at Columbia University and went on to earn his Ph.D. in astronomy from the University of California at Berkeley in 1973. Before taking up his present job he held appointments at University College London and the University of California at Los Angeles. He is currently an Alfred P. Sloan Foundation research fellow. In 1981 he was awarded the Newton Lacey Pierce Prize of the American Astronomical Society. Margon's main research interests are extrasolar X-ray astronomy, ultraviolet astronomy and observations of the sources of X rays by means of visible radiation.

HUA JUE-MING ("The Mass Production of Iron Castings in Ancient China") is associate research fellow at the Institute of the History of Natural Science of the Chinese Academy of Sciences in Beijing. He studied mechanical engineering at Qinghua University, where he was graduated in 1958. From 1956 until 1964 he did work on the history of smelting and casting in China under Lin Xian-zhou and Zhang Zi-gao. He then moved to the Institute of the History of Natural Science as a graduate student specializing in the history of metalworking. He is a council member of the Chinese Society of the History of Science and Technology and head of the society's committee on the history of metallurgical technique. Hua's recent work has been on the casting of Shang-Zhou bronzes, the construction of pre–Qing Dynasty bells and the cast iron of the Han and Wei dynasties.

SUPERIOR STANDARDS IN MOLECULAR BIOLOGY AND BIOCHEMISTRY

Second Edition

Biochemistry

Lubert Stryer, *Stanford University*

Systematically developing each topic from the chemical principle to its biological and medicinal applications, the Second Edition of *Biochemistry* is organized around five major themes: conformation and dynamics; generation and storage of metabolic energy; biosynthesis of macromolecular precursors; information; and molecular physiology. The relationship between three-dimensional structure and biological activity is exemplified throughout in exceptionally clear, five-color illustrations. With many important additions — including treatment of 150 new topics, over 250 new illustrations, and a wealth of material demonstrating the power of biochemistry in elucidating physiological mechanisms — the Second Edition of *Biochemistry* is the definitive text for standard first courses in the field.

"[T]his is an excellent textbook for the general biochemist as well as for the medically oriented reader.... Though it is primarily a superb introduction, even experts will profit from its perspective and clarity of outlook."
—The New England Journal of Medicine

1981, 949 pages, 983 illustrations
hardbound: ISBN 0-7167-1226-1 $32.95

DNA Replication

Arthur Kornberg, *Stanford University*

DNA Replication presents thorough coverage of the numerous DNA transactions that determine the structure and function of genetic material. Written by a Nobel laureate in medicine, this comprehensive study is augmented by illustrative tables and figures throughout, and provides a definitive account of knowledge and research through 1979.

"[T]his book is outstanding. Its concise, eloquent style makes it a delight to read, while the numerous excellent figures and tables and the comprehensive lists of references provide the reader with a wealth of relevant information."*—SEARCH*

1980, 724 pages, 308 illustrations
hardbound: ISBN 0-7167-1102-8 $38.95

1982 Supplement to DNA Replication

Arthur Kornberg, *Stanford University*

An essential addition to the study of DNA biochemistry, *1982 Supplement to DNA Replication* contains new figures, tables, and a review of important literature and developments in the field since 1979. The organization of topics in the *Supplement* parallels that of *DNA Replication*. Designed for use either independently or as a companion volume to Kornberg's earlier work, *1982 Supplement to DNA Replication* is indispensable reading for researchers and students in this rapidly developing field.

1982, 273 pages, 48 illustrations
paper: ISBN 0-7167-1410-8 $10.00

Second Edition

Physical Biochemistry
Applications to Biochemistry and Molecular Biology

David M. Freifelder, *Brandeis University*

A comprehensive introduction to important techniques of modern biochemistry and molecular biology, the Second Edition of *Physical Biochemistry* emphasizes the types of data yielded by these techniques and shows how these data can be interpreted. Theory is presented intuitively, and all problems have a strong experimental orientation. The text has been extensively revised and expanded to incorporate new developments in the field. Among the features of the Second Edition are an expanded section on macromolecules, presentation of the latest developments in autoradiography, inclusion of twice as many problems as in the First Edition, and treatment of numerous new subjects.

1982, 761 pages, 573 illustrations
hardbound: ISBN 0-7167-1315-2 $39.95
paper: ISBN 0-7167-1444-2 $21.95

Molecules to Living Cells
Readings from *Scientific American*

With Introductions by
Philip C. Hanawalt, *Stanford University*

Spanning the past four decades, this collection of twenty-six articles from *Scientific American* by distinguished scientists emphasizes molecular structures, their interactions in forming supramolecular structures, and the genetic control of their synthesis in the growing cells that constitute life. The book includes introductions providing background information, updating earlier articles, and placing scientific developments in perspective.

1980, 340 pages, 280 illustrations
hardbound: ISBN 0-7167-1208-3 $22.00
paper: ISBN 0-7167-1209-1 $10.95

To order copies of any of the above books, send a check for the total amount (please include $1.50 to cover postage and handling; California residents add appropriate sales tax) to: Promotion Department, Box MAG, W. H. Freeman and Company, 660 Market Street, San Francisco, CA 94104.

METAMAGICAL THEMAS

Virus-like sentences and self-replicating structures

by Douglas R. Hofstadter

Two years ago, when I first wrote about self-referential sentences in this department, I was hit by an avalanche of mail from readers intrigued by the phenomenon of self-reference in its many different guises. I had the chance to print some of those responses a year ago, and that column triggered a second wave of responses. Many of them have illuminated self-reference in new ways. Here I should like to describe the ideas of several people, beginning with two who responded to my initial column with remarkably similar letters: Stephen Walton of New York and Donald R. Going of Oxon Hill, Md.

Walton and Going saw self-replicating sentences as being similar to viruses: small objects that enslave larger and more self-sufficient "host" objects, getting the hosts by hook or by crook to carry out a complex sequence of replicating operations that bring into existence new copies of the virus, which are then free to go off and enslave further hosts. "Viral sentences," as Walton called them, are "those that seek to obtain their own reproduction by commandeering the facilities of more complex entities."

Both Walton and Going were struck by the perniciousness of such sentences: the selfish way they invade a space of ideas and manage, merely by making copies of themselves, to take over a large portion of that space. Why do they not manage to overrun *all* of the space? It is a good question. The answer should be obvious to students of evolution: the sentences do not do so because of competition from other self-replicators. One type of replicator seizes one region of the space and becomes good at fending off rivals; thus a "niche" in idea space is carved out.

This idea of an evolutionary struggle for survival by self-replicating ideas is not original with Walton or Going, although both of them had fresh things to say about it. The first reference I know of to this notion is in a passage by the neurophysiologist R. W. Sperry in an article he wrote in 1965 titled "Mind, Brain, and Humanist Values." He writes: "Ideas cause ideas and help evolve new ideas. They interact with each other and with other mental forces in the same brain, in neighboring brains, and, thanks to global communication, in far distant, foreign brains. And they also interact with the external surroundings to produce in toto a burstwise advance in evolution that is far beyond anything to hit the evolutionary scene yet, including the emergence of the living cell."

Shortly thereafter, in 1970, the molecular biologist Jacques Monod came out with his book *Chance and Necessity*. In its last chapter he wrote of the selection of ideas as follows:

"For a biologist it is tempting to draw a parallel between the evolution of ideas and that of the biosphere. For while the abstract kingdom stands at a yet greater distance above the biosphere than the latter does above the nonliving universe, ideas have retained some of the properties of organisms. Like them, they tend to perpetuate their structure and to breed; they too can fuse, recombine, segregate their content; indeed they too can evolve, and in this evolution selection must surely play an important role. I shall not hazard a theory of the selection of ideas. But one may at least try to define some of the principal factors involved in it. This selection must necessarily operate at two levels: that of the mind itself and that of performance.

"The performance value of an idea depends upon the change it brings to the behavior of the person or the group that adopts it. The human group upon which a given idea confers greater cohesiveness, greater ambition, and greater self-confidence thereby receives from it an added power to expand which will insure the promotion of the idea itself. Its capacity to 'take,' the extent to which it can be 'put over' has little to do with the amount of objective truth the idea may contain. The important thing about the stout armature a religious ideology constitutes for a society is not what goes into its structure, but the fact that this structure is accepted, that it gains sway. So one cannot well separate such an idea's power to spread from its power to perform.

"The 'spreading power'—the infectivity, as it were—of ideas is much more difficult to analyze. Let us say that it depends upon preëxisting structures in the mind, among them ideas already implanted by culture, but also undoubtedly upon certain innate structures which we are hard put to identify. What is very plain, however, is that the ideas having the highest invading potential are those that *explain* man by assigning him his place in an immanent destiny, in whose bosom his anxiety dissolves."

Monod refers to the universe of ideas, or what I earlier called idea space, as "the abstract kingdom." Since he portrays it as a close analogue to the biosphere, we could as well call it the "ideosphere."

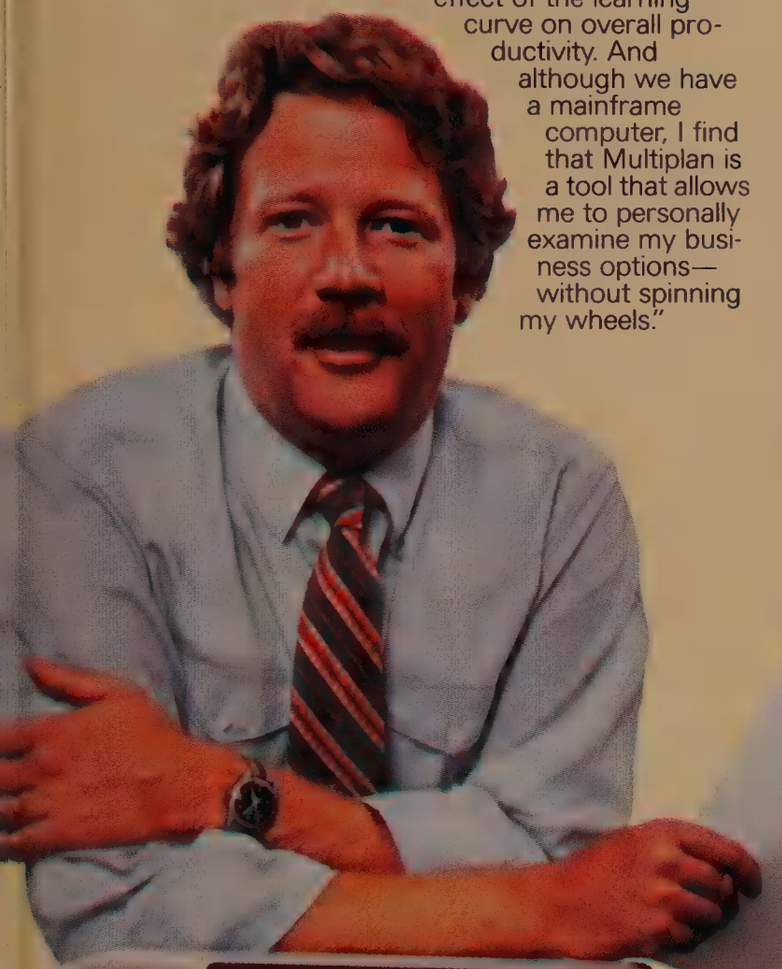
In 1976 the evolutionary biologist Richard Dawkins published his book *The Selfish Gene*, whose last chapter further develops this theme. Dawkins' name for the unit of replication and selection in the ideosphere—the ideosphere's counterpart to the biosphere's gene—is "meme." Just as a library is an organized collection of books, so a memory is an organized collection of memes. And the soup in which memes grow and flourish—the analogue to the "primordial soup" out of which life first oozed—is the soup of human culture. Dawkins writes:

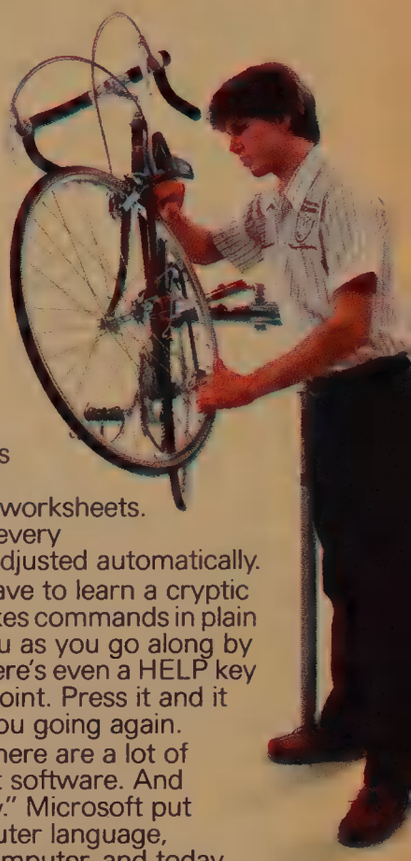
"Examples of memes are tunes, ideas, catch-phrases, clothes fashions, ways of making pots or of building arches. Just as genes propagate themselves in the gene pool by leaping from body to body via sperms or eggs, so memes propagate themselves in the meme pool by leaping from brain to brain via a process which, in the broad sense, can be called imitation. If a scientist hears, or reads about, a good idea, he passes it on to his colleagues and students. He mentions it in his articles and his lectures. If the idea catches on, it can be said to propagate itself, spreading from brain to brain. As my colleague N. K. Humphrey neatly summed up an earlier draft of this chapter: '...memes should be regarded as living structures, not just metaphorically but technically. When you plant a fertile meme in my mind, you literally parasitize my brain, turning it into a vehicle for the meme's propagation in just the way that a virus may parasitize the genetic mechanism of a host cell. And this isn't just a way of talking—the meme for, say, "belief in life after death" is actually realized physically,

"Microsoft® Multiplan™ allows me to explore more alternatives in less time. I think it leads to better management decisions."

Edward R. Schwinn, Jr.
President, Schwinn Bicycle Company

"Multiplan software helps me make better business decisions. It's as simple as that. With Multiplan on my microcomputer, I can explore a number of alternatives faster. That's really important to me in planning and tracking my totally hand-crafted Schwinn Paramount Bicycle. As with anything handmade, cost control and product planning are major considerations. For instance, Multiplan lets me study the effect of the learning curve on overall productivity. And although we have a mainframe computer, I find that Multiplan is a tool that allows me to personally examine my business options—without spinning my wheels."





A tool for decision makers. Microsoft Multiplan is a microcomputer software program that can help you too. In many ways. You can analyze cash flow. Plan budgets. Forecast income. Manage production. Multiplan allows you to set up an electronic worksheet for whatever your need may be. And, it lets you explore alternatives faster because when you make one change you immediately see the results of that change throughout the worksheet and on any related worksheets. Just change one number and every number that depends on it is adjusted automatically.

In plain English. You won't have to learn a cryptic language to use Multiplan. It takes commands in plain English. And, it will prompt you as you go along by telling you what to do next. There's even a HELP key in case you need help at any point. Press it and it gives you information to get you going again.

Backed by our knowhow. There are a lot of good reasons to use Microsoft software. And among the best is "knowhow." Microsoft put BASIC, the first microcomputer language, on the very first microcomputer, and today, Microsoft software is running on well over a million computers. Worldwide.

See it in action. If you'd like to explore more alternatives yourself, ask your local computer dealer for a demonstration of Microsoft Multiplan. Multiplan is available for most popular personal computers, including Apple,® the IBM® PC and computers that use CP/M-80® or MS™-DOS operating systems. It's a good management tool for building bicycles. It's also a good management tool for building businesses. Like yours.

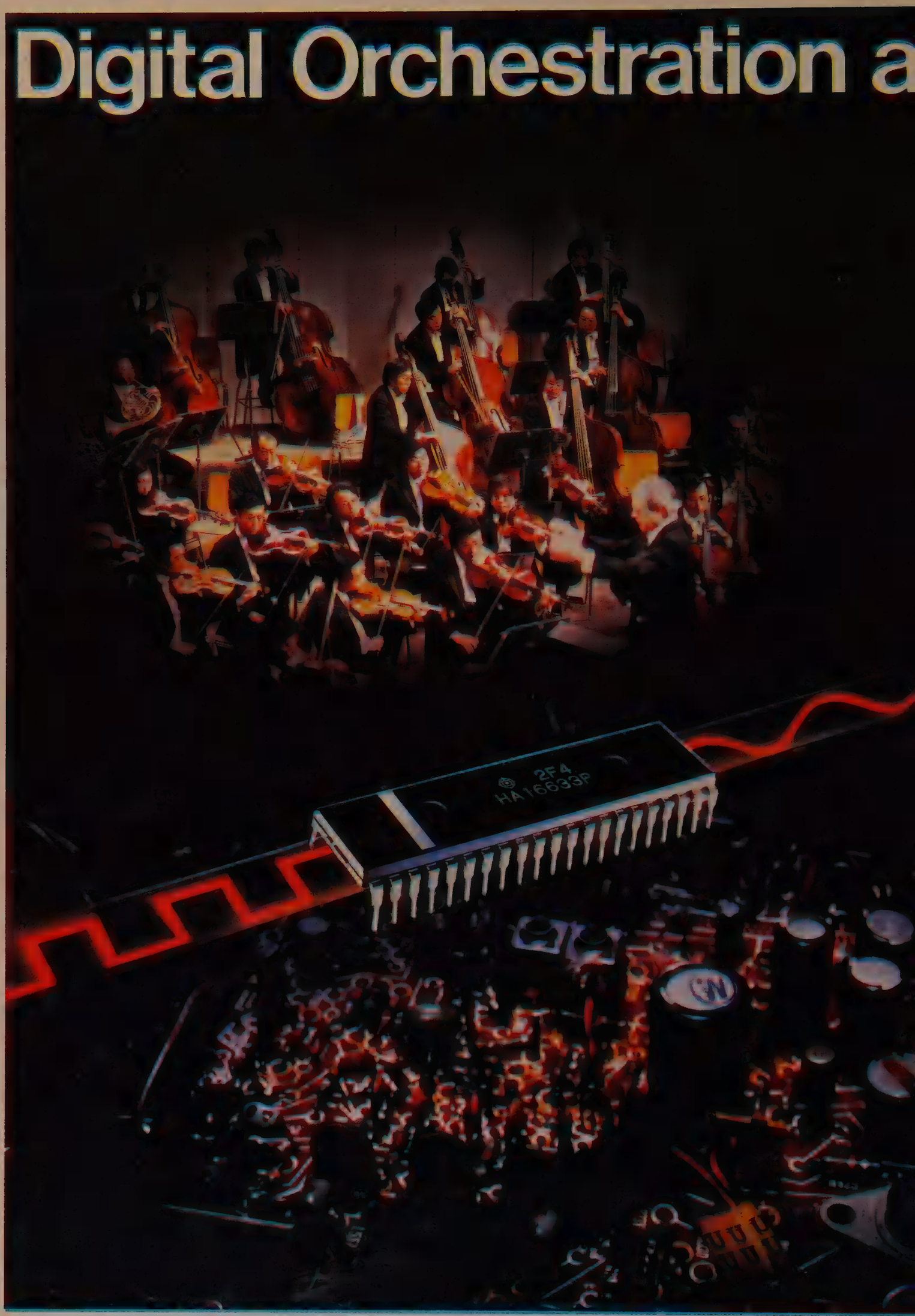
BETTER TOOLS FOR MICROCOMPUTERS

MICROSOFT.

MICROSOFT CORPORATION, 10700 NORTHUP WAY, BELLEVUE, WASHINGTON 98004

Microsoft is a registered trademark, and Multiplan, MS, and the Microsoft logo are trademarks of Microsoft Corporation. Apple is a registered trademark of Apple Computer, Inc. IBM is a registered trademark of International Business Machines, Corp. CP/M-80 is a registered trademark of Digital Research, Inc.

Digital Orchestration a



the Hands of a Master

Many can play; few can conduct

The latest innovation in audio, the digital audio disc (DAD) player, is a complex package of circuitry which would fill an entire filing cabinet if conventional transistors were used. Large-scale integrated circuits (LSIs) are the obvious way to conserve space and achieve electronic efficiency. But few audio component makers have the technological resources necessary to design and manufacture their own LSIs. That's one reason why DAD players have not been made commercially available. Until now.

Hitachi leads the way with integrated circuitry

Hitachi was able to draw on its own advanced semiconductor capabilities, developed in the course of serving the computer industry, in order to design the DAD player's nine different LSIs. For each LSI, a circuit of discrete components equivalent to some 32,000 transistors had to be assembled in a large rack. Following testing, the circuit was drawn on paper. Then, the drawing was photographically reduced and printed on a silicon wafer – the heart of each LSI. At Hitachi's factories, these original LSIs can be mass-produced at a 99.9% flaw-free rate.

Coordinating the overall performance

One of the most important LSIs in the player is the 16-bit linear digital/analog converter. It's responsible for translating digital pickup signals into analog signals, processing 44,100 sound segments per second, for playback through standard amplifier/speaker systems.

Other key LSIs in the player include those to control the laser tracking mechanism, a data strobe IC, data and control metal-oxide semiconductor LSIs, a 16-bit RAM, a 4-bit microcomputer for system control, and two sample hold ICs for left and right channel output. All together, they form an extremely compact, solid-state system – perfectly orchestrated because it's the product of a single, integrated electronics manufacturer.

Hitachi's DA 1000 digital audio disc player

Technical excellence is embodied in all Hitachi products

The development of LSIs for the DAD player is just one case demonstrating Hitachi's technological strength. You'll find other examples in all of Hitachi's products, from microprocessors and bubble memory devices to a whole host of high-quality electronic appliances. Our comprehensive technical expertise is your guarantee of convenience, easy operation and high reliability in every product that bears the Hitachi brand.

HITACHI

A World Leader in Technology

millions of times over, as a structure in the nervous systems of individual men the world over.'

"Consider the idea of God. We do not know how it arose in the meme pool. Probably it originated many times by independent 'mutation.' In any case, it is very old indeed. How does it replicate itself? By the spoken and written word, aided by great music and great art. Why does it have such high survival value? Remember that 'survival value' here does not mean value for a gene in a gene pool, but value for a meme in a meme pool. The question really means: What is it about the idea of a god which gives it its stability and penetrance in the cultural environment? The survival value of the god meme in the meme pool results from its great psychological appeal. It provides a superficially plausible answer to deep and troubling questions about existence. It suggests that injustices in this world may be rectified in the next. The 'everlasting arms' hold out a cushion against our own inadequacies which, like a doctor's placebo, is none the less effective for being imaginary. These are some of the reasons why the idea of God is copied so readily by successive generations of individual brains. God exists, if only in the form of a meme with high survival value, or infective power, in the environment provided by human culture."

Dawkins takes care here to emphasize that there need not be an exact copy of each meme, written in some universal memetic code, in each person's brain. Memes, like genes, are susceptible to variation or distortion—the analogue of mutation. Various mutations of a meme will have to compete with one another, as well as with other memes, for attention, that is, for brain resources in terms of both space and time devoted to that meme. Memes must compete not only for inner resources but also, since they are transmissible visually and aurally, for radio and television time, billboard space, newspaper and magazine column-inches and library shelf space. Furthermore, some memes will tend to discredit others, and some groups of memes will tend to be internally self-reinforcing. Dawkins writes:

"Mutually suitable teeth, claws, guts, and sense organs evolved in carnivore gene pools, while a different stable set of characteristics emerged from herbivore gene pools. Does anything analogous occur in meme pools? Has the god meme, say, become associated with any other particular memes, and does this association assist the survival of each of the participating memes? Perhaps we could regard an organized church, with its architecture, rituals, laws, music, art, and written tradition, as a co-adapted stable set of mutually assisting memes.

"To take a particular example, an aspect of doctrine which has been very effective in enforcing religious observance is the threat of hell fire. Many children and even some adults believe that they will suffer ghastly torments after death if they do not obey the priestly rules. This is a particularly nasty technique of persuasion, causing great psychological anguish throughout the Middle Ages and even today. But it is highly effective. It might almost have been planned deliberately by a Machiavellian priesthood trained in deep psychological indoctrination techniques. However, I doubt if the priests were that clever. Much more probably, unconscious memes have ensured their own survival by virtue of those same qualities of pseudo-ruthlessness which successful genes display. The idea of hell fire is, quite simply, *self-perpetuating,* because of its own deep psychological impact. It has become linked with the god meme because the two reinforce each other, and assist each other's survival in the meme pool.

"Another member of the religious meme complex is called faith. It means blind trust, in the absence of evidence, even in the teeth of evidence.... Nothing is more lethal for certain kinds of meme than a tendency to look for evidence.... The meme for blind faith secures its own perpetuation by the simple unconscious expedient of discouraging rational inquiry.

"Blind faith can justify anything. If a man believes in a different god, or even if he uses a different ritual for worshipping the same god, blind faith can decree that he should die—on the cross, at the stake, skewered on a Crusader's sword, shot in a Beirut street, or blown up in a bar in Belfast. Memes for blind faith have their own ruthless ways of propagating themselves. This is true of patriotic and political as well as religious blind faith."

When I muse about memes, I often find myself picturing an ephemeral flickering pattern of sparks leaping from brain to brain, shouting "Me, me!" Walton's and Going's letters reinforced this image in interesting ways. For instance, Walton begins with the simplest imaginable viral sentences—"Say me" and "Copy me"—and moves quickly to more complex variations with blandishments ("If you copy me, I'll grant you three wishes") or threats ("Say me or I'll put a curse on you"), neither of which, he observes, is likely to be able to keep its word. Of course, as he points out, this may not matter, the only final test of viability being success at survival in the meme pool. All's fair in love and war—and war includes the eternal battle for survival, in the ideosphere no less than in the biosphere.

To be sure, very few people above the age of five will fall for the simpleminded threats or promises of these sentences. If, however, you simply tack on the phrase "in the afterlife," far more people will be lured into the memetic trap. Walton observes that a similar gimmick is used by the typical chain letter (or "viral text"), which "promises wealth to those who faithfully replicate it and threatens doom to any who fail to copy it." Do you remember the first time you received such a chain letter? Do you remember the sad tale of "Don Elliott, who received $50,000 but then lost it because he broke the chain"? And the grim tale of "General Welch in the Philippines, who lost his life six days after he received this letter because he failed to circulate the prayer—but before he died, he received $775,000"? It is hard not to be just a little sucked in by such tales, even if you wind up throwing the letter out contemptuously.

I found Walton's phrases "viral sentence" and "viral text" to be exceedingly catchy—little memes in themselves, definitely worthy, in my opinion, of replication some 700,000 times in print, and who knows how many times orally beyond that? Walton's own viral text, as you can see before your eyes, has managed to commandeer the facilities of a powerful host: an entire printing press and magazine and distribution service. It has leaped aboard and is now—even as you read this viral sentence—propagating itself madly through the ideosphere.

This idea of choosing the right host is itself an important aspect of the quality of a viral entity. Walton puts it this way: "The recipient of a viral text can, of course, make a big difference. A tobacco mosaic virus that attacks a salt crystal is out of luck, and some people rip up chain letters on sight. A manuscript sent to an editor may be considered viral, even though it contains no explicit self-reference, because it is attempting to secure its own reproduction through an appropriate host; the same manuscript sent to someone who has nothing to do with publishing may have no viral quality at all."

As Walton's letter ends it graciously steps forward from the page and squeaks to me directly on its own behalf: "Finally, I (this text) would be delighted to be included, in whole or in part, in your next discussion of self-reference. With that in mind, please allow me to apologize in advance for infecting you."

Whereas Walton mentioned Dawkins in his letter, Going seems not to have been aware of Dawkins at all, which makes his letter quite remarkable in its close connection to Dawkins' ideas. Go-

ing suggests that we consider, to begin with, Sentence *A:* "It is your duty to convince others that this sentence is true." As he says: "If you were foolish enough to believe this sentence, you would attempt to convince your friends that *A* is true. If they were equally foolish, they would convince their friends, and so on until every human mind contained a copy of *A* in it. Thus *A* is a self-replicating sentence. More particularly, it is the intellectual equivalent of a virus. If Sentence *A* were to enter a mind, it would take control of the mind's machinery and exploit it to produce hundreds of copies of itself in other minds.

"The problem with Sentence *A,* of course, is that it is absurd; no one could possibly believe it. However, consider the following:

System *S:*
Begin:
*S*1: Blah.
*S*2: Blah blah.
*S*3: Blah blah blah.
: :
: :
: :
*S*99: Blah blah blah blah blah blah....
*S*100: It is your duty to convince others that System *S* is true.
End.

Here *S*1 through *S*99 are meant to be statements that taken together constitute a belief system having some degree of coherency. If System *S* taken as a whole were convincing, then the entire system would be self-replicating. System *S* would be particularly convincing if *S*100 were not stated explicitly but held as a logical consequence of the other ideas in the system."

Let us refer to Going's *S*100 as the "hook" of System *S,* for it is by this hook that System *S* hopes to hoist itself onto a higher level of power. Note that on its own a hook that says in effect, "It is your duty to believe me," is not a viable viral entity; in order to "fly" it needs to drag something extra along with it, just as a kite needs a tail to stabilize it. Pure lift goes out of control and self-destructs, but controlled lift can lift itself along with its controller. Similarly, *S*100 and *S*1–*S*99 (taken as a set) are symbiotes: they play complementary, mutually supportive roles in the survival of the meme they constitute together. Now Going develops the theme a little further:

"Are there any real idea systems that behave like System *S?* I know of at least two. Consider the following:

System *X:*
Begin:
*X*1: Anyone who does not believe System *X* will burn in hell.
*X*2: It is your duty to save others from suffering.
End.

"If you believed in System *X,* you would attempt to save others from hell by convincing them that System *X* is true. Thus System *X* has an implicit 'hook' that follows from its two explicit sentences, and so System *X* is a self-replicating idea system. Without being impious, one may suggest that this mechanism has played some small role in the spread of Christianity.

"Note that System *X* is still not credible. It takes quite a bit of skill and luck to produce a believable idea system. A more sinister form of self-replication is Sentence *B:* 'The bourgeoisie is oppressing the proletariat.' If you believed *B,* you would want to liberate the proletariat from the bourgeoisie. You would quickly discover that you were not strong enough for the purpose. You would need the help of thousands of like-minded people. The first step in obtaining their help would be to convince them that *B* is true. Thus a 'hook' follows from Sentence *B,* and *B* is therefore a self-replicating idea system.

"Statement *B* is merely a special case of the generalized statement, Sentence *V:* 'The *villain* is *wronging* the *victim.*' Here the word *villain* must be replaced with the name of some real group (capitalists, communists, imperialists, Jews, freemasons, aristocrats, men, foreigners, etc.). *Victim* must be replaced with the corresponding victim and *wronging* filled in as desired. The result will be a self-replicating idea system for the same reasons as *B* was. Note that each of the suggested substitutions yields a historically attested idea system. In fact, it is difficult to think of a form of political extremism that does not reduce to *V* and therefore to the 'hook,' *S*100. One hesitates to explain real historical events in terms of such a silly mechanism, and yet...."

Going brings his ideas to an amusing conclusion as follows: "Suppose we parody my thesis by proposing Sentence *E:* 'The self-replicating ideas are conspiring to enslave our minds.' This 'paranoid' statement is clearly an idea of type *V.* Thus the thesis seems to describe itself. Further, if we accept *E,* then we must say that this type-*V* idea implies that we must distrust all ideas of type *V.* This is the Epimenides paradox."

It is interesting that all these people who have explored these ideas have given examples ranging from the very small scale of such things as catchy tunes (Dawkins cites the opening theme of Beethoven's Fifth Symphony) and phrases (the word "meme" itself) to the very large scale of ideologies and religions. Dawkins uses the term "meme complex" for these larger agglomerations of memes; I prefer the single word "scheme."

One reason is that it fits well with the usage suggested by the psychiatrist and writer Allen Wheelis in his novel *The Scheme of Things.* Its central character is a psychiatrist and writer named Oliver Thompson, whose darkly brooding essays are scattered throughout the book, interspersed with brightly colored, evocative episodes. Thompson is obsessed with the difference between, on the one hand, "the raw nature of existence, unadorned, unmediated," which he refers to repeatedly as "the way things are," and, on the other hand, "schemes of things" invented by human beings—ways of making order and sense out of the way things are. Here are some of Thompson's musings on that theme:

"I want to write a book... the story of one man whose life becomes a metaphor for the entire experience of man on earth. It will portray his search through a succession of schemes of things, show the breakdown, one after another, of each pattern he finds, his going on always to another, always in the hope that the scheme of things he finds and for the moment is serving is *not* a scheme of things at all but reality, the way things are, therefore an absolute that will endure forever, within which he can serve, to which he can contribute, and through which he can give his mortal life meaning and so achieve eternal life....

"The scheme of things is a system of order. Beginning as our view of the world, it finally *becomes* our world. We live within the space defined by its coordinates. It is self-evidently true, is accepted so naturally and automatically that one is not aware of an act of acceptance having taken place. It comes with one's mother's milk, is chanted in school, proclaimed from the White House, insinuated by television, validated at Harvard. Like the air we breathe, the scheme of things disappears, becomes simply reality, the way things are. It is the lie necessary to life. The world as it exists beyond that scheme becomes vague, irrelevant, largely unperceived, finally nonexistent....

"No scheme of things has ever been both coextensive with the way things are and also true to the way things are. All schemes of things involve limitation and denial....

"A scheme of things is a plan for salvation. How well it works will depend upon its scope and authority. If it is small, even great achievement in its service does little to dispel death. A

scheme of things may be as large as Christianity or as small as the Alameda County Bowling League. We seek the largest possible scheme of things, not in a reaching out for truth, but because the more comprehensive the scheme the greater its promise of banishing dread. If we can make our lives mean something in a cosmic scheme, we will live in the certainty of immortality. Those attributes of a scheme of things that determine its durability and success are its scope, the opportunity it offers for participation and contribution, and the conviction with which it is held as self-evidently true. The very great success of Christianity for a thousand years follows upon its having been of universal scope, including and accounting for everything, assigning to all things a proper place; offering to every man, whether prince or beggar, savant or fool, the privilege of working in the Lord's vineyard; and being accepted as true throughout the Western world.

"As a scheme of things is modified by inroads from outlying existence, it loses authority, is less able to banish dread; its adherents fall away. Eventually it fades, exists only in history, becomes quaint or primitive, becomes, finally, a myth. What we know as legends were once blueprints of reality. The Church was right to stop Galileo; activities such as his import into the regnant scheme of things new being which will eventually destroy that scheme."

Taken in Wheelis' way, "scheme" seems a fitting replacement for Dawkins' "meme complex." A scheme imposes a top-down kind of perceptual order on the world, propagating itself ruthlessly, like Going's System *S* with its 'hook.' Wheelis' description of the inadequacy of all "schemes of things" to fully and accurately capture "the way things are" is strongly reminiscent of the vulnerability of all sufficiently powerful formal systems to either incompleteness or inconsistency—a vulnerability that follows from another kind of 'hook': the famous Gödelian hook, which arises from the capacity for self-reference of such systems, although neither Wheelis nor Thompson makes any mention of the analogy. We shall come back to Gödel shortly.

The reader of *The Scheme of Things* must be struck by the professional similarity between Wheelis and his protagonist. It is impossible to read the book and not surmise that Thompson's views are reflecting Wheelis' own—and yet who can say? It is a tease. Even more tantalizing is the title of Thompson's book, which Wheelis casually mentions toward the end of the novel: it is *The Way Things Are*—a striking contrast to the title of the book in which it is contained. One wonders: What is the meaning of this elegant literary pleat, in which one level folds back on another? What is the symbolism of Wheelis within Wheelis?

Such a twist, by which a thing (a sentence, a book, a system, a person) seems to refer to itself but does so only by allusion to something *resembling* itself, is called "indirect self-reference." You can do this by pointing at your image in a mirror and saying, "That person sure is good-looking!" This one is very simple, because the connection between something and its mirror image is so familiar and obvious-seeming to us that there seems to be no distance whatever between direct and indirect referents: we equate them completely. Hence it seems there is no referential indirectness.

On the other hand, this depends on the ease with which our perceptual systems convert a mirror image into its reverse, and on other qualities of our cognitive systems that enable us to see through several layers of translation without being aware of the layers—like looking through many feet of water and seeing not the water but only what lies at the bottom.

Some indirect self-references are of course subtler than others. Consider the case of Sam and Sue, a couple ostensibly having a conversation about their friends Tina and Tim. It happens that Sam and Sue are having some problems in their relationship, and those problems are quite analogous to those of Tina and Tim, only with the sexes reversed: Sam is to Sue what Tina is to Tim in their respective relationships. Therefore as Sam and Sue's conversation progresses, although on the surface it is about their friends Tina and Tim, on another level it is *actually* about themselves as they are reflected in these other people. It is almost as if by talking about Tina and Tim, Sam and Sue are going over a fable by Aesop that has obvious relevance to their own plight. There are things going on simultaneously on two levels, and it is hard to tell how conscious either of the participants is of the exchange of dual messages—one of concern about their friends, one of concern about themselves.

Indirect self-reference can be exploited in the most unexpected and serious ways. Consider the case of President Reagan, who on a recent occasion of high Russian-American tension over Iran went out of his way to recall President Truman's behavior in 1945, when Truman made some very blunt threats to the Russians about the possibility of the U.S.'s using nuclear weapons if necessary against any Russian move in Iran. Merely by bringing up the memory of that occasion, Reagan was inviting a mapping to be made between himself and Truman and thereby was issuing a not-so-veiled threat, although no one could point to anything explicit. There simply was no way that a sufficiently conscious being could fail to make the connection. The resemblance between the two situations was too blatant.

Does self-reference really come, then, in two varieties—direct and indirect—or are the two types just distant points on a continuum? I would say unhesitatingly that it is the latter. Furthermore, you can delete the prefix "self-," so that the question becomes one of reference in general. The essence is simply that one thing refers to another whenever, to a conscious being, there is a sufficiently compelling mapping between the roles the two things are perceived to play in some larger structures or systems. Caution is needed here. By "conscious being" I mean an analogy-hungry perceiving machine that gets along in the world thanks to its perceptions; it need not be human or even organic. Actually I would carry the abstraction of the term "reference" even further, as follows. The mapping of systems and roles that establishes reference need not actually be *perceived* by any such being; it suffices for the mapping to exist and simply be *perceptible* to such a being.

The movie *The French Lieutenant's Woman* (made from John Fowles's novel of the same name) provides an elegant example of ambiguous degrees of reference. It consists of interlaced vignettes from two concurrently developing stories both of which involve complex romances; one takes place in Victorian England, the other in the present. The fact that there are two romances already suggests, even if only slightly, that a mapping is called for. But much more than that is suggested. There are structural similarities between the two romances: each of them has triangular qualities, and each story focuses on only one leg of the triangle. Moreover, the same two actors play the lovers in both romances, so that you see them in alternating contexts and with alternating personality traits. The reason for this "coincidence" is that the contemporary story concerns the making of a film of the Victorian story.

As the two stories unfold in parallel a number of coincidences arise that suggest even more strongly that a mapping should be made. It is left, however, to the viewer to carry out this mapping; it is never explicitly called for. After a time, though, it simply becomes unavoidable. What is pleasant in this game is the fluidity left to the viewer: there is much room for artistic license in seeing connections, or suspecting or even inventing connections.

Indirect reference of the artistic type

is much less precise than indirect reference of the formal type. The latter arises when two formal systems are isomorphic, that is, when they have strictly analogous internal structures, so that there is a rigorous one-to-one mapping between the roles in the one and the roles in the other. In such a case the existence of genuine reference becomes as clear to us as it is in the case of someone talking about his or her mirror image: we take it as an immediate, pure self-reference, without even noticing the indirectness, the translational steps mediated by the isomorphism. In fact, the connection may seem too direct even to be called "reference"; some may see it simply as identity.

This perceptual immediacy is the reason a famous statement of mathematical logic, Gödel's sentence *G*, is called self-referential. Everyone accepts the idea that *G* talks about a *number g;* the tricky Gödelian step is in seeing that the number *g* plays a role in the system of natural numbers strictly analogous to the role the *sentence G* plays in the axiomatic system it is expressed in. This Wheelis-like oblique reference by *G* to itself by way of its "image" *g* is generally accepted as being genuine self-reference. (Note that we have even a further mapping: *G* plays the role of Wheelis, and its Gödel number *g* that of Wheelis' alter ego Thompson.)

The two abstract mappings that establish *G*'s self-reference but make it seem indirect can be collapsed into just one mapping, following a slogan we might formulate this way: "If *A* refers to *B*, and *B* is just like *C*, then *A* refers to *C*." For instance, we can let *A* and *C* be Wheelis, with *B* representing Thompson. This makes Wheelis' self-reference a "theorem." Of course, this "theorem" is not rigorously proved, since our slogan has to be taken with a grain of salt. Being "just like" something else is a highly disputable matter.

In a formal context where "is just like" is virtually synonymous with "plays a role isomorphic to that of," however, the slogan can have a strict meaning, and thereby it can justify a theorem more rigorously. In particular, if *A* and *C* are equated with *G*, and *B* is equated with *g*, then our slogan runs: "If *G* refers to *g*, and *g* plays a role isomorphic to that of *G*, then *G* refers to *G*." Since the premises are true, the conclusion must be true. According to this scheme of things, then, *G* is a genuinely self-referential sentence rather than some kind of logical illusion as deceptive as an Escher print.

Indirect self-reference suggests the idea of indirect self-replication, in which a viral entity, instead of replicating itself exactly, brings into being another entity that plays the same role as it does but in some other system: perhaps its translation into French, perhaps a string of the product numbers of all its parts, together with preaddressed envelopes containing checks made out to the factories where those parts are made, and a list of instructions telling what to do with all the parts when they arrive in the mail.

That may sound familiar to some readers. In fact, it is an indirect reference to the Von Neumann Challenge, the puzzle I posed in this department last January to create a self-describing sentence whose only quoted matter is at the word or letter level rather than at the level of entire quoted phrases. I discovered, as I received candidate solutions, that many readers did not understand what the requirement meant. The challenge came out of an objection to the complexity of the "seed" (the quoted part) in W. V. Quine's version of the Epimenides paradox:

"yields falsehood when appended to its quotation" yields falsehood when appended to its quotation.

To see what is strange here, imagine that you want to have a space-roving robot build a copy of itself out of raw materials it encounters in its travels. One way you could do it would be to make the robot symmetrical, like a human being. You could also make the robot able to make a mirror-image copy of any structure it encounters along its way. Finally, have the robot programmed to scan the world constantly, the way a hawk scans the ground for prey. The search image in the robot's case is that of an object identical with its own left half. The robot need not be *aware* that its target is identical with its left half; the search can go on merrily for what seems to it to be merely a very complex and arbitrary structure. When, after scouring the universe for 17 googolplex years, it finally comes across such a structure, then of course the robot activates its mirror-image-production facility and creates a right half. The last step is to fasten the two halves together, and presto, a copy emerges. It is as easy as pie—provided you are willing to wait 17 googolplex years (give or take a few minutes).

The arbitrary and peculiar aspect of the Quine sentence, then, is that its seed is half as complex (which is to say nearly as complex) as the sentence itself. If we resume our robot parable, what we should ideally like in a self-replicating robot is the ability to make itself literally from the ground up: let us say, for instance, to mine iron ore, to smelt it, to cast it, to make nuts and bolts and sheet metal out of it and so on, and finally to be able to assemble the small parts into progressively larger subunits until a replica is born out of truly raw materials. That was the spirit of the Von Neumann Challenge: I wanted a linguistic counterpart to the "self-replicating robot of the second kind."

In particular, this means a self-documenting or self-building sentence that builds both its halves—its quoted seed and its unquoted building rule—out of linguistic raw materials (words or letters). The mistake many readers made was to present as the seed a long sequence of individually quoted words (or letters) in a specific order, then to exploit that order in the building rule. They might as well have quoted one big long ordered string, as Quine did. The idea of my challenge, in contrast, was that all structure in the built object must arise exclusively out of some principle enunciated in the building rule, not out of the seed's internal structure.

Just as a self-replicating robot in some random alien environment is hardly likely to find all its parts lined up on a shelf in order of assembly but must rely on its "brain" or program to recognize raw parts wherever and whenever they turn up so that it can grab them and therefrom assemble a copy of itself, so the desired sentence must treat the pieces of the seed without regard to the order in which they are listed and yet be able to construct itself in the proper order out of them. Thus it is fine if you enclose the entire seed within a single pair of quotation marks rather than quoting each word individually—all that matters is that the word order of the seed (or better still the letter order) not be exploited.

The seed of the ideal solution would be a long inventory of parts, similar to the list of ingredients of a recipe—perhaps a list of 50 *e*'s, then 46 *t*'s, and so on. Clearly those letters cannot remain in that order; they simply constitute the raw materials out of which the new sentence is to be built. No one sent in a solution whose seed was at such a primordial level. A few people, however, did send in adequate, if not wonderfully elegant, solutions with seeds at the word level. The first correct solution I received came from Frank Palmer of Chicago, who therefore gets the first "Johnnie" award—a self-replicating dollar bill that is given every other decade to the Grand Winner of the Von Neumann Challenge. Unfortunately the dollar bill consumes the entire body of its owner in its bizarre process of self-replication, and so it is wisest to simply lock up the bill to protect oneself from its voracious appetite.

Palmer submitted several versions. In them he utilized uppercase and lower-

case to distinguish between the seed and the building rule respectively. Here is one solution, slightly modified by me:

after alphabetizing, decapitalize FOR AFTER WORDS STRING FINALLY UPPERCASE FIGPBVKXQJZ NONSENSICAL DECAPITALIZE SUBSTITUTING ALPHABETIZING, finally for nonsensical string substituting uppercase words

Let us watch how it works, step by careful step. We must bear in mind that the instructions we are following are the lowercase words printed above and that the uppercase words are not to be read as instructions. Nor, for that matter, are the lowercase words we shall soon be working with. They are like the inert, anesthetized body of a patient being operated on, who, when the operation is over, will wake up and become animate. So let's go. First we are to alphabetize the seed. (I am treating the comma as attached to the word it follows.) This gives us the following:

AFTER ALPHABETIZING, DECAPITALIZE FIGPBVKXQJZ FINALLY FOR NONSENSICAL STRING SUBSTITUTING UPPERCASE WORDS

Next we are to decapitalize it. This will yield some lowercase words—the "anesthetized" lowercase words I mentioned above:

after alphabetizing, decapitalize figpbvkxqjz finally for nonsensical string substituting uppercase words

All right; now our final instruction is to locate a nonsensical string (that's easy: "figpbvkxqjz") and substitute for it the uppercase words (namely a copy of the original seed itself). This yields:

after alphabetizing, decapitalize FOR AFTER WORDS STRING FINALLY UPPERCASE FIGPBVKXQJZ NONSENSICAL DECAPITALIZE SUBSTITUTING ALPHABETIZING, finally for nonsensical string substituting uppercase words

And this is a perfect copy of our starting sentence! It can now wake up from its anesthesia and replicate itself in turn.

The critical step was the first one: alphabetization. This turns the arbitrarily ordered seed into a grammatical, meaningful sentence—merely by mechanically exploiting a presumed knowledge of the ABC's. But why not? It is perfectly reasonable to presume superficial typographical knowledge about letters and words, since such knowledge deals with printed material as raw material: purely syntactically, without regard to the meanings carried therein. This is just like the way enzymes in the living cell deal with the DNA and RNA they chop up and alter and piece together again: purely chemically, without regard to the "meanings" carried therein. Just as chemical valences and affinities and so on are taken as givens in the workings of the cell, so alphabetical and typographical facts are taken as givens in the Von Neumann Challenge.

In Palmer's solution the seed happens to have been ordered by the length of words, but that is inessential; any random order would have done, and that is the crucial point many readers missed. Another rather elegant solution was sent by Martin Weichert of Munich. It runs this way:

Alphabetize and copy in quotes "in and copy quotes Alphabetize"

It works on the same principle as Palmer's sentence and again features a seed whose internal structure is irrelevant to the self-replication. Weichert also sent a palindromic solution in Esperanto, in which the flexible word order of the language plays a key role. Michael Borowitz and Bob Stein of Durham, N.C., sent a solution similar to Palmer's.

Finally, last year's gold-medal winner for self-documentation, Lee Sallows, was a bit piqued by my suggestion that the gold on his medal was somewhat tarnished since he had not paid close enough attention to the use-mention distinction. Apparently I goaded him into constructing an even more elaborate self-documenting sentence, reproduced at the left. Although it is not self-replicating, since it does not spell out its own construction explicitly, it is another marvelous Sallowsian gem, and I shall allow the gold on his medal to go untarnished this year. (I hope this will satisfy those purists who wrote to me insisting that gold does not tarnish!)

```
                              *
                          Write
                      down ten 'a's,
                  eight 'c's, ten 'd's,
            fifty-two 'e's, thirty-eight 'f's,
        sixteen 'g's, thirty 'h's, forty-eight 'i's,
    six 'l's, four 'm's, thirty-two 'n's, forty-four 'o's,
   four 'p's, four 'q's, forty-two 'r's, eighty-four 's's,
seventy-six 't's, twenty-eight 'u's, four 'v's, four 'W's,
 eighteen 'w's, fourteen 'x's, thirty-two 'y's, four ':'s,
       four '*'s, twenty-six '-'s, fifty-eight ','s,
             sixty '('s and sixty ')'s, in a
                  palindromic sequence
                      whose second
                        half runs
                          thus:
                          :suht
                        snur flah
                      dnoces esohw
                   ecneuqes cimordnilap
              a ni ,s')' ytxis dna s'(' ytxis
        ,s',' thgie-ytfif ,s'-' xis-ytnewt ,s'*' ruof
   ,s':' ruof ,s'y' owt-ytriht ,s'x' neetruof ,s'w' neethgie
  ,s'W' ruof ,s'v' ruof ,s'u' thgie-ytnewt ,s't' xis-ytneves
    ,s's' ruof-ythgie ,s'r' owt-ytrof ,s'q' ruof ,s'p' ruof
    ,s'o' ruof-ytrof ,s'n' owt-ytriht ,s'm' ruof ,s'l' xis
         ,s'i' thgie-ytrof ,s'h' ytriht ,s'g' neetxis
              ,s'f' thgie-ytriht ,s'e' owt-ytfif
                   ,s'd' net ,s'c' thgie
                      ,s'a' net nwod
                          etirW
                            *
```

Lee Sallows' self-documenting sentence

BOOKS

The sky of the year, natural fires, cacti, city-states and mitochondria

by Philip Morrison

1983 GRAPHIC TIMETABLE OF THE HEAVENS, The Maryland Academy of Sciences and Scientia, Inc. Scientia, Inc., 1815 Landrake Road, Baltimore, Md. 21204 ($6.50). The night sky spreads before the eye in just two dimensions, giving only hints of depth. Those who would guide us to look the right way generally map that vault on a flat page and sample the flow of the third dimension—time—by offering a sequence of maps. The distortions of flattening projections and the use of dots for the luminaries, whether black on white or the somewhat more natural reversal to white symbols on a night black ground, add a measure of abstraction that taxes the rank beginner, although it can soon be mastered.

This colorful wall poster is by comparison highly abstract. It spreads the night sky for the entire year across one single glossy surface the size of four magazine pages. The sky is shrunk effectively to one dimension, along the zodiacal arc where the action is: the path of sun, moon and planets. Down the central axis of the page run the midnights, with each point down the page representing one day later among the 365 days of the year. Plotted across the page, the hours of each night roll by. An hourglasslike profile marks out the times of sunset and sunrise, the long nights at top and bottom, the shorter summer nights in midyear.

Sinuously winding to and fro across sunset and sunrise frontiers, two curves, one orange and one green, display the rising and setting of Mercury and of Venus on each day of the year; two other paths in color, only subtly curved away from simple straight lines, show the outer planets. Mars wends its more ambiguous way, one year as regular as a superior planet, the next year curving strongly as the planet passes behind the sun. A few stars are straight slanting lines. This careful geocentric representation, timed by the sun underfoot, presents the visible appearances in an abstract pattern as epicyclic as Ptolemy's. Not for the entirely uninitiated, the poster is both handsome and practical for those who know the way to their seats in the celestial theater but want a careful program.

The curves are ingeniously marked and repeated to allow easy interpolation for a wide range of latitudes, although the plot is direct to an accuracy of one minute of time (with refraction and parallax corrections) for a location in Illinois. Even finicky users anywhere from Los Angeles to Toronto will have no trouble. Moon phases and the times of moonrise and moonset, with planetary and solar standstills and conjunctions, are also to be read off the curve-filled sheet.

After the group performance of 1982, the quieter year ahead offers no spectacles to adorn the stately procession. The eclipse seasons are well marked; two lunar eclipses this year are "easy if not somewhat bland" in these parts, neither of them total. (The total solar eclipse of next midsummer sends the cognoscenti to study maps of Java.) Bright Venus enters conjunction with Mars twice in autumn morning skies. Shy Mercury should be well placed for view in the late-April evening twilight.

This reliable and calligraphic chart is no new conception; it has been widely admired for a generation. Only lately, however, has it found a format so strong and colorful.

FIRE IN AMERICA: A CULTURAL HISTORY OF WILDLAND AND RURAL FIRE, by Stephen J. Pyne. Princeton University Press ($35). Fire across the land is not as ancient as ocean shore or as oxygen in the atmosphere, but it does go back to the time when plants first learned well the trick of living on dry land, a span a hundred times our hominid one. Thus fire in the New World did not need to wait for the coming of the first hunters but burned freely wherever it was ignited by high-current lightning discharge in a dry season. Down to our own day lightning is a chief predator of trees and grass; perhaps a tenth of all wildland fires are started by it.

Fire is a species with a defined range; the Tropics are too wet to harbor much of it and icy lands lack fuel to feed it. In the broad Temperate Zone it is at home, with good growth piling up many seasons' fuel on forest floor and across rolling prairie, until drought might make the fiery release of that stored energy fully self-propagating. The venerable sequoia is adapted to withstand fire by its thick asbestoslike bark; longevity demands such preparation. An oxygenated world heaped with dried cellulose is patently metastable. Flame is to be seen as a rapid mode of inevitable organic decay, an essential if stochastic part of catabolism.

Nature's fires in North America were multiplied by the hunters from Asia. Our Grandfather Fire served the people from the eastern woodlands to the lands of the Paiute and the Modoc. Broadcast fire in Indian hands was a weapon, a purifier, a key technique of hunting and a sovereign recycler, able to maintain both the light yields of the gatherers of nut and acorn and the higher yields of the woodland growers of maize. The grasslands that dominate the landscape of the U.S. are in large part the artifact of millenniums of purposeful firing by the Indian population. The forest primeval is in the main a literary device; it was widely cleared away before ever Europeans arrived with their iron: ax, gun and will.

Fire is easily tolerable to a society of nomads, people already on the move. Settlers in log cabins or in canyon houses above Malibu fear it; industrial societies see accessible timber as potential wealth, yearn for wilderness and adventure and are able to muster mass human energy and powerful technical weapons against wildfire. Above all, a countryside long cleared by ax, plow and graze calls for reforestation as successful farmlands shift.

The great barbecue of the Gilded Age had profligately consumed the nation's resources. It was in the U.S. Geological Survey that the idea of natural-resource management first crystallized a century ago. John Wesley Powell's concern was water for the dry West. Powell himself saw virtue in the ancient uses of fire: "The best thing to do for the Rocky Mountain forests was to burn them down." After about 1910, however, the old ways came to appear all but treasonous; the heroism of the firefighter joined with the romantic appeal of the wilderness to support the total exclusion of wildfire by moral sanctions beyond any mere prudent husbandry. The Chief Forester over the decades became more and more of a fire chief. From Theodore Roosevelt's day to 1930 the Forest Service fought its fires at the frontier, in economic defense of transitional values. For the next 40 years it supported a policy of strong control: "Each succeeding day will be planned... with an aim of obtaining control before ten o'clock of the next morning." Since 1970 the service has come once again to view fire by prescription as an admissible part of the processes that rule every forest.

The young Iowa historian who writes

Tina has never had a Teddy Bear.

A mother's love. A doll to cuddle. Tina knows nothing of these things. But she does know fear, rejection, and hunger.

For just $18 a month, you can help save a child like Tina.

Through our sponsorship program you can help provide a child with a better diet, clothes, medical attention, school. And even a toy or two.

But don't wait. There are so many. And somewhere, right now, a child is dying from starvation and neglect.

Write to: Mrs. Jeanne Clarke Wood Children, Incorporated, P. O. Box 5381, Dept. SA1T3, Richmond, Va. 23220 USA

☐ I wish to sponsor a boy ☐, girl ☐, in ☐ Asia, ☐ Latin America, ☐ Middle East, ☐ Africa, ☐ USA, ☐ Greatest Need.
☐ I will give $18 a month ($216 a year). Enclosed is my gift for a full year ☐, the first month ☐. Please send me the child's name, story, address and picture.
☐ I can't sponsor, but will help $_____
☐ Please send me further information.
☐ If for a group, please specify.

_____
Church, Class, Club, School, Business, etc.

NAME _____

ADDRESS _____

CITY _____ STATE _____ ZIP _____

U.S. gifts are fully tax deductible.
Annual financial statements are available on request.

CHILDREN, INC.

this book is as concerned with the ebb and flow of the metaphors behind policy as he is with the documented infighting of a passel of Federal bureaus. One brilliant chapter treats of fire and warfare. Fire as weapon is very old, but as cultures grew over the land the indiscriminate nature of broadcast fire spoke against resorting to it. Open fire was increasingly confined to unequal conflicts between indigenous populations resisting enclosure or relocation and their better-armed rulers and occupiers. Even the descendants of Europeans dwelling in the piny woods of the U.S. Southeast used forest fires to defend traditional hunting grounds and pasture against the newer industrial forestry.

In war between industrial nations open fire became rare. "The images of... war are those of mud, not of fire." World War II rekindled airborne offensive wildfire; "an old word was revived to describe a new weapon: firestorm." In the years from 1950 to 1970 the military issue began to remold the U.S. Forest Service. Smokey the Bear had evolved directly out of a wartime advertising campaign based on Bambi, when the rights to the Disney design became too expensive. Aircraft, smoke jumpers, helicopters and Army crews entered the fire line. Financial and research links to war preparedness grew strong. Twenty years ago the Forest Service research laboratories examined controlling the "greatest fire on earth," the consequence of two imagined 50-megaton explosions over Portland and Seattle.

By the 1970's a new shift had come. Military research was terminated; urban fires were to be studied elsewhere than in Forest Service laboratories; even the received ecological theory changed from a determinist view of green forest, the natural crown of a sequence of stages, to a probabilistic cycling of subtle rhythms that might in part be powered by fire. Fire was again natural, and prescribed fire in the forest was once again imperative. Now it is the exclusion of fire that needs justification; it appears as a naive act of transient intervention. Has this periodic shift of concerns ended? It might be noted that the summer of 1982 saw a serious study of global thermonuclear war (in the Swedish journal *Ambio*) concluding that the most threatening of worldwide consequences would be particulates from the mass fires, engendering a hemispheric darkness at noon for at least one growing season, hence a terrible famine.

Professor Pyne writes with wit and profundity. A reader must prepare for a sometimes disturbing juxtaposition of original and even universal perceptions with the tendentious doings of Federal officeholders of less than great distinction. Pyne himself was crew foreman with the North Rim Longshots amid burning pines along the Grand Canyon for many seasons; his book did not arise out of a merely intellectual inquiry. The detailed treatment of nine geographic regions as case histories enriches the volume beyond a scale that can be covered in a review.

Our symbiotic pact with fire is ancient indeed, pervasive in language and myth. Surely it goes back to the old artificers, *Homo erectus*. For a long time fire was elemental, metaphysical, a principle of explanation. But as the poet-wit John Donne saw in the Galilean years, "The Element of fire is quite put out." It became only one gas-phase reaction, a molecular mechanism in grate and cylinder. To take its place there has arisen a more abstract new element, energy, the grand generalization of fire's power. Only in its scale is wildfire still elemental. The great burns transcend all civil releases of energy; they rage at the yield of megatons per day, like earthquakes. The most extensive fires all but enter on the gigaton domain, the potential of the stocked warheads worldwide. Philosopher–fire fighter that he is, Pyne is an optimist: "If our ancestors, with little more than river cobbles and chipped flint, could seize fire, surely modern mankind can cope with its own versions of the forbidden flame."

THE CACTI OF THE UNITED STATES AND CANADA, by Lyman Benson. With line drawings by Lucretia Breazeale Hamilton. Stanford University Press ($85). "At the point where we left the Gila, there stands a Cereus six feet in circumference and so high, I could not reach half way to the top of it with the point of my sabre by many feet." So wrote the young officer who led a reconnaissance along the new border from Fort Leavenworth to San Diego in 1846, the first account of the giant saguaro cactus to reach print. That very grove of giants is still there; how they look is suggested, with a certain understandable hyperbole, in the grand engraving reproduced from the full report of that trip, which presented to "a waiting world" its first view of the Arizona desert.

Professor Benson has produced a splendid volume to carry to the public the results of his 48 years of study of that remarkable plant family, the New World succulents. It is an exemplary volume, accessible, rich in scientific and historical context, profusely and thoughtfully illustrated, short on jargon and yet meticulous with the documentation obligatory for the taxonomic botanist. The treatment of record and meaning is so fastidious as to inspire any veteran of the Philadelphia bar.

There are in effect three books within the handsome 1,000-page volume. The first is a general account of the cacti, their structure, physiology and ecology along with many taxonomic issues. One chapter, whose aim is to clarify the geo-

graphic distribution and associations of the cacti (there are native prickly pears on Cape Cod, and cacti are found wild in 45 U.S. states and five provinces of Canada), admirably describes and illustrates floristic associations from the foggy Aleutians to the Florida Keys. The bulk of the text is given over to a modern and expert flora of the cacti, each genus keyed, described and mapped in detail with frequent careful drawings and photographs showing details of growth and structure; 48 color plates display the bizarre forms and lovely flowers typical of these forms.

Some 60 pages in fine print constitute a third book of documentation for readers who might enjoy the minutiae of taxonomic recognition. Professor Benson is particularly careful to describe the evidence, from the correct use and preparation of herbarium sheets to the character of devoted and extensive botanizing in the field. One map dots out the counties this professional has visited in his intricate life path among the cacti; another page remarks that one tiny but elegant species has cost more man-days of search than any other, save a still smaller one. "Finding the first one required many days of searching bent over or on hands and knees." The saguaro grove along the Gila was visited by the author in the course of his own journey to all the 1846 campsites to gather specimens of the cacti first described and sketched, but not collected, on that survey 120 years earlier.

Students of the family "have been preoccupied with discovery of new entities, real or otherwise." Professor Benson chooses rather to follow the intellectual lead of the St. Louis physician George Engelmann, whose unmatched genius for organizing botanical information illumines his pioneer work on cacti. An excess of discrimination mars most subsequent classification of cacti, with their diversity magnified by the industrious gardeners. This book treats of cacti in nature, part of the wild flora of the region studied.

The subject is too detailed to summarize, but a few results may serve as a sample for the general reader. Cacti are wax-coated plants, stems almost without leaves, low in surface for their volume (tinier forms retract entirely into the ground on hot days), spiny to safeguard their water stores from animal attack. Those spines are in fact the vestige of leaves; recent experiments on one form of prickly pear showed that the same buds could develop into either leaves or spines. The structures treated with one familiar plant hormone turned into normal leaves, those treated with another became spines.

Carbon dioxide is unable to diffuse through the waxy cuticle of the cactus; it is admitted by open stomata, in the cool darkness when the loss of water through

MERGER OR FINANCING?

Are you thinking of equity financing or merger of your company now or in the future? Beckman maintains a long-term search for top-quality technical growth opportunities serving medicine, science and industry. In addition to our traditional fields of instrumentation and precision electronic components, increased emphasis is being focused on clinical-diagnostic reagents, applied molecular biology and computer software technology. Perhaps you or an associate would like a copy of our Diversification Interests statement — for your present review or future reference. Available on request and in confidence.

ROBERT B. BROWN
Director of Corporate Planning

BECKMAN

Beckman Instruments, Inc.
2500 Harbor Blvd., Fullerton, CA 92634 USA

AUDIO-FORUM® offers the best in self-instructional foreign language courses using audio cassettes — featuring those used to train U.S. State Dept. personnel in Spanish, French, German, Japanese, Greek, Arabic, Chinese, Italian, and more.

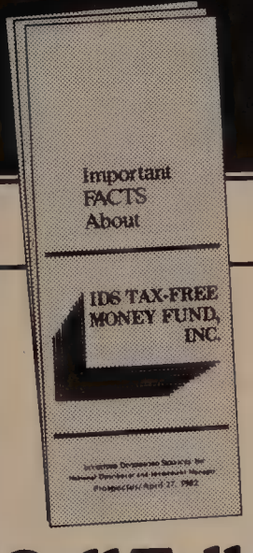
Learn a foreign language on your own! Free catalog.

Call (203) 453-9794, or fill out and send this ad to—
Audio-Forum
Suite 51-C On-the-Green
Guilford, CT. 06437

Name
Address
City
State/Zip
I am particularly interested in (check choice):
☐ Spanish ☐ French ☐ German
☐ Greek ☐ Russian ☐ Swahili
☐ Other

The IDS Tax-Free Money Fund

...FREE Consumers' Guide To This Federally Tax-Exempt Money Market Fund.



Call or Write for Your No-Cost, No-Obligation Kit and Get . . .

■ A Fact Book on The IDS Tax-Free Money Fund . . . fully explains how a tax-free* money market fund can pay today's higher, 100% Federally tax-exempt yields—without any sales charges or "early withdrawal penalties."
■ An Informative Prospectus of The Fund's Many Advantages . . . explaining free drafts for "checkwriting" access, the goal of a Constant $1 Share Price and more.
■ An Easy-to-Read Investment Profile Report . . . showing the latest quotation of tax-free yield, a Profile of The Fund's high-quality securities . . . and Much More.

*A portion may be subject to some state and local taxes.

Call Toll-Free 1-800-IDS-IDEA

. . . That's TOLL-FREE 1-800-437-4332

IDS, IDS Tower, P.O. Box 369, Minneapolis, MN 55440

For more complete information, including management fees and expenses, write or call for a prospectus. Read it carefully before you invest or send money.

Name
Address City
State Zip Telephone

IDS Since 1894

03-376-144-001

the opening is low. At night photosynthesis cannot fix the essential gas, and so it is stored at low temperature bonded into a simple organic acid, ready to be utilized by daylight. Thus cactus juices are a sour morning refreshment; that barrel cactus may be juicy, but its water is acceptable only in an emergency. The saguaro flesh has an unpleasant flavor, but the copious fruit of the right prickly pear is a major food in the markets of highland Mexico. Burbank's spineless cactus wonder turned out to be one more public-relations hype, less his fault than that of the swindlers who exploited his name in hilarious controversy during the years before World War I. The arresting purples that sometimes spread over the spiny green pads are recorded in one color photograph. They are unique to desert scenes; certain pigments formed in cacti (and in one related order) are chemically distinct from the anthocyanins of all other flowering plants.

The windowsill horticulturist is not the intended reader of this book; its cost and bulk are disproportionate to that interest. More important, the treatise is carefully limited in geographic range, aimed not at the collector's array of succulents but at the cacti there on the American land. For those who seek out waste places, and for apropos libraries of all degrees, it is a model of devotion, learning and thoughtful execution.

One final warning emerges: The cacti are the most threatened of all plant families. Collection and sale of cactus seeds—always in great surplus—are to be encouraged, but living plants ought not to be removed from the field either for sale or for the home. New, small protected areas are chosen for special reasons; a "single fantastic soil outcrop" supports one rare species over a range 20 miles long and only a quarter of a mile wide. On Indian land, that plant has owners who are both aware and protective. The danger list, however, is some 70 species long.

THE CITY-STATE IN FIVE CULTURES, edited by Robert Griffith and Carol G. Thomas. ABC-Clio, Inc., Rivera Campus, Santa Barbara, Calif. 93103 ($22.50). Most of these cities evoke images at once: Uruk, Sparta, Giotto's Florence, Kepler's Weil der Stadt. Fewer readers can quickly call up the massive walls of Kano or Zaria, the Hausaland examples of this genus of social organization. In an unusually readable volume five scholar-specialists from the shores of Puget Sound present overviews of the history and nature of comparable city-states. The oldest of all cities were the dozen temple towns of Sumer 5,000 years ago; the latest city taken up here lasted until 1812, when the Sword of Truth, 'Uthman dan Fodio, set his own sons over the Hausa towns as viceroys of a new Fulani empire across the Nigerian grasslands.

The aim is compelling; it is to crystallize by detailed comparison a single entity out of these diverse chronicles. In Sumer the people lived by painstaking irrigation of the barley fields, in Florence by the export of wool cloth intricately finished, in the medieval cities of Germany and the Swiss cantons by the artisanship of the guildsmen and the capital stores of bankers and merchants. The wide Hausa town walls protectively encircled fertile gardens, unfailing water holes and even ironstone hills for the smiths. No single economy unites these narratives. Our taxonomists claim no overarching or final system for the perceived unity; they present modestly enough an inductive view.

A city-state first of all is mononuclear. A single urbanized core is its heart, bounded by some membrane: a wall or a moat. That nucleus must be nourished, and so some hinterland is organic to the city, which always seeks economic autonomy within its own small region, more often finding its livelihood only by forming the node of a vigorous system of trade and finance. No city-state is a nation; usually the cities that are its allies and rivals are bearers of the same language and culture. A world outside seems given over to barbarians. The Hausa legend accounted for the kingdoms that surrounded their lands as states founded by the descendants of their own originator, Abuyazidu, although only through the sons of his concubine: the bastard Seven. The final decisive item of the key is independence: the city-state is the single sovereign of its citizens, who come to share a certain kinship, if only a fictitious one.

The sample is not small. There were more than 1,500 Greek *poleis,* and dozens or scores of city-states in all the other cultures considered. They share scale at least in order of magnitude: in space, say a few square miles within the walls; in head count, 10,000 urban families, not including the peasant folk outside; in time, half a thousand years of proud independence. None of the city-states persist. They were all caught in tension, the need to expand control as local resources declined. Or a new perception of potential arose, against the imperative of a consolidated, thriving and devoted core. Frank slavery, selfish rule by a hundred noble families (the Florentines tore down the arrogant towers of the nobility when once Florence looked like San Gimignano today) and the loss of well-worn lucrative routes of trade all stressed the city-states. Their weakness, like their strength, was in their scale. They were poorest at adapting to sweeping structured change. A wider unity finally swallowed them all.

City-states seem to represent one stable form of social grouping when central authority is lacking. Striking, convergent in nature along a spectrum of culture, history and economics, they were neither inevitable stages nor enduring structures as human polity evolved. They were conspicuous eddies in the turbulent flow.

MITOCHONDRIA, by Alexander Tzagoloff. Plenum Press ($42.50). A century ago cytologists, tireless microscopists adept at the fixing and staining of tissue, saw small, threadlike granules (*mitochondria* in the Greek synthesis) scattered by the score within the cells of every plant and animal. They resembled bacteria in size and form. In the first years after World War II it was demonstrated that these intricate organelles held a special importance. The classical biochemistry of solutions, armed with the irresistible Waring blender and an arsenal of chemical and photochemical weapons, had worked out step by delicate step the long chains of metabolism; the crown of them all is the Krebs cycle. There eight enzymes cooperate in a wheel of linked reactions. The first product of organic oxidation is eventually burned to water and carbon dioxide, at last releasing to the air-breathing cells of higher life nearly 20 times as much energy from each glucose molecule as anaerobic fermentation can yield.

In 1948 it became clear that the entire system was housed in the mitochondria. The energy released by the Krebs cycle, however, must somehow be stored for biochemical work. That sequence, so to speak a cog meshed with the Krebs cycle, yields the universal cellular fuel, ATP, by a series of oxidative pathways that overall store chemically in the unstable phosphate all the energy won from glucose. The process lay outside the grasp of the classical methods; its key enzymes are insoluble, and cell-free extracts are still poor producers of ATP. Not mere protein solutions but cellular ultrastructure holds the intricate mechanism of ATP synthesis.

This crisp text by a Columbia participant of the discipline makes its point at once to the eye. Revealing electron micrographs of mitochondria go back 30 years, to the pioneering of Albert Claude and George E. Palade. The microscopists of our electronic times have their own cunning means of fixation and staining that resolve structure at a fineness a couple of hundred times smaller than the optics of an earlier generation could. Mitochondria repay such close scrutiny; indeed, it has turned out that the bag within a bag that is a granule does hold all the enzymes of the Krebs process in its innermost compartment. The outer membrane wall is more or less rod-shaped, but the inner one, seen in writhing motion in rare cinema shots of living examples, is complex, deeply reentrant, changing.

Thermal Conduction Module

IBM

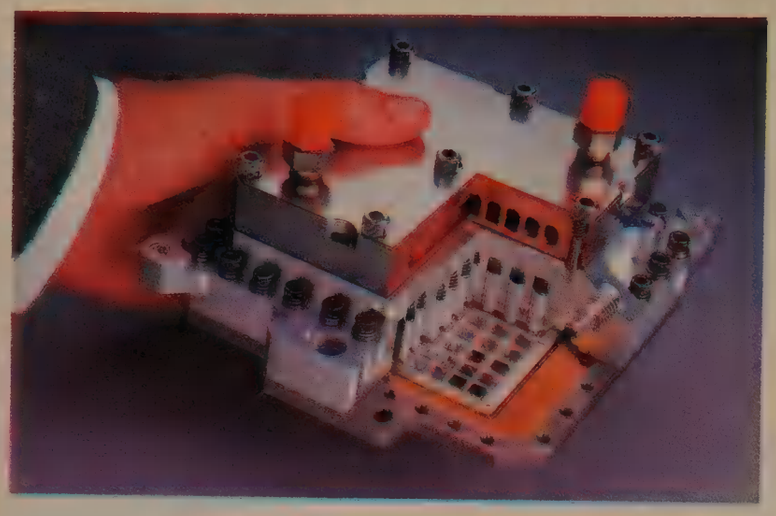
To: Jim
From: Bill
Subject: IBM Technology

I've been reviewing some of our past and present technological achievements, and it occurred to me that the scientific, engineering and academic communities might like to know more about them. Will you select one from the following list as the first topic? Thanks.

Vacuum tube digital multiplier
IBM 603/604 calculators
Selective Sequence Electronic Calculator (SSEC)
Tape drive vacuum column
Naval Ordnance Research Calculator (NORC)
Input/output channel
IBM 608 transistor calculator
FORTRAN
RAMAC and disks
First automated transistor production
Chain and train printers
Input/Output Control System (IOCS)
STRETCH computer
"Selectric" typewriter
SABRE airline reservation system
Removable disk pack
Virtual machine concept
Hypertape

System/360 compatible family
Operating System/360
Solid Logic Technology
System/360 Model 67/Time-Sharing System
One-transistor memory cell
Cache memory
Relational data base
First all-monolithic main memory
Thin film recording head
Floppy disk
Tape group code recording
Systems Network Architecture
Federal cryptographic standard
Laser/electrophotographic printer
First 64K-bit chip mass production
First E-beam direct-write chip production
Thermal Conduction Module
288K-bit memory chip
Robotic control language

Bill-
Good idea! Our TCM is a state-of-the-art technology.
Let's start there.
Jim



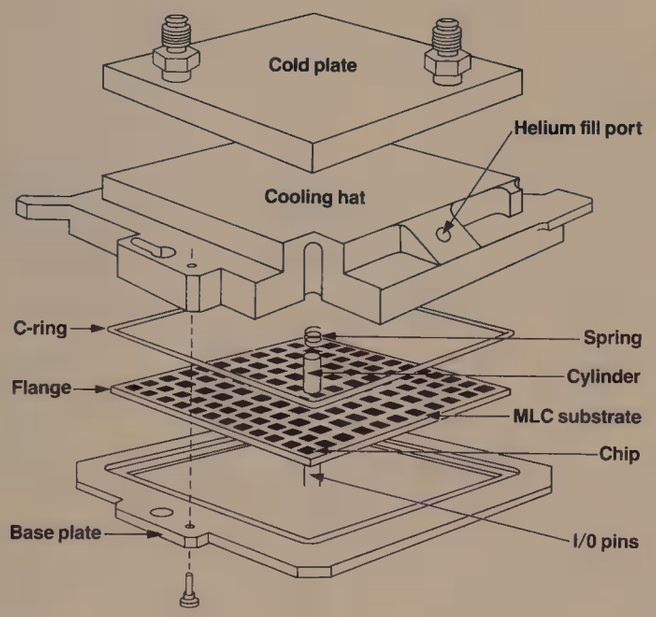
Cutaway of TCM (15x15x6 cm) which contains logic and buffer memory equivalent to an IBM System/370 Model 148. Its new packaging, cooling and assembly technology provides IBM's largest computers with the greatest circuit density yet reported.

The time it takes a signal to travel between circuits is key to a computer's performance. The shorter the distance the signal has to travel, the faster the computer can operate. Shortening the distance requires improving the circuit density of the packaging and interconnections, as well as placing the circuits closer together on the logic and memory chips.

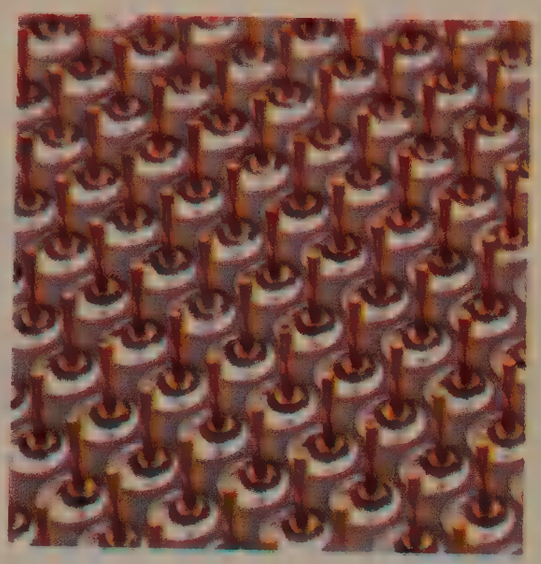
Circuit packaging in IBM's most powerful computers—the 3081, 3083 and 3084—is the densest yet reported in the industry. This has been made possible by our Thermal Conduction Module (TCM) and the way it is combined with a high-performance circuit board.

A TCM holds up to 133 logic and memory array chips on a 90-mm substrate, consisting of 33 layered ceramic sheets. More than 120 meters of wiring within these layers connect up to 45,000 circuits within the TCM. And 1,800 pins are brazed to the substrate bottom to plug the module into a circuit board.

All of this requires computer-aided



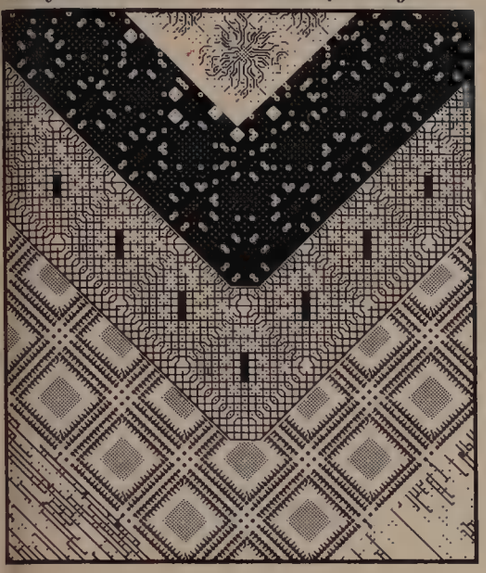
Exploded view of TCM assembly.



Detail of input/output pins brazed to back of multilayer ceramic (MLC) substrate which plug TCM to next level of packaging.

Thermal Conduction Module

Magnified portion of ceramic sheets (0.2 or 0.28 mm thick, unfired), the basic building blocks of the 33-layer MLC. A typical substrate contains 350,000 vias for layer-to-layer connections and 130 m of wiring.



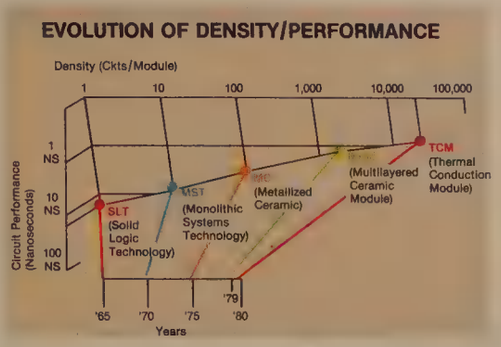
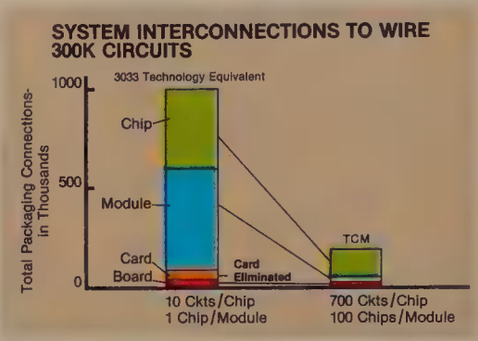


Illustration of TCM's dramatic benefits. Average performance improved more than 20x over 1965 level. Average circuits/module increased from one in SLT to 30,000 in TCM.



Effect of TCM technology on hypothetical 300K circuit system. Circuits/chip and chips/module increased with dramatic reductions of system interconnections and elimination of card-level package. Result is lowered cost, improved performance and reliability.

design techniques and precision manufacturing operations unique to IBM.

Thus "packaged," the chips require a new cooling mechanism to remove the heat created by their unprecedented density. Within the hermetically-sealed TCM, spring-loaded cylinders contact the chips and conduct heat away from them. Helium surrounds the chips and cylinders, providing an inert atmosphere. Its heat-conducting efficiency is approximately six times that of air, and it carries the heat to an aluminum alloy "hat" atop the TCM. The heat is then conducted to a cold plate through which water circulates at 24°C.

The circuit density of the TCM required IBM engineers to develop a high density, high-performance printed-circuit board on which to mount the modules and to provide power and signal paths. This board is the densest yet reported for large-scale manufacturing.

This 600 x 700 mm printed-circuit board contains 20 layers of circuitry, and includes more than one kilometer of wiring. It can accommodate up to six or nine TCMs depending on the board configuration.

Together, the TCM and the circuit board eliminate an entire level of packaging previously required in computers. This combination contains all the functions normally attributed to modules, cards, boards and interconnecting cables in prior technologies.

The circuit density made possible by this new packaging, cooling and assembly technology allows a single TCM to contain as much logic and buffer memory as an IBM System/370 Model 148. Today's large-scale IBM 3081 carries up to 26 TCMs on four circuit boards which contain its complete logic—nearly 800,000 circuits—and displace only four cubic feet.

IBM scientists and engineers worldwide have contributed technologies to the TCM. These technologies are all part of our continuing commitment to research and development, funded with more than $8 billion over the past seven years.

IBM®

This 3081 printed-circuit board carries nine TCM's and is, in itself, a major advance in circuit packaging.

A recent issue of the IBM Journal of Research and Development is devoted to the IBM 3081. For a free copy, write: Director of Technical Publications, IBM Corporation, Dept. 100, 44 S. Broadway, White Plains, N.Y. 10601.

This spring the Lindblad Explorer visits Melanesia, Malaysia, Burma. You come, too.

The lands of gold. This spring Salén Lindblad's Explorer takes you to their perfumed magic. Three incredible cruises starting March 4th.

All in total comfort aboard the **Lindblad Explorer*** in outside cabins and featuring superb Scandinavian and native food. All accompanied by experts who will help you learn about these lands of gold. And about yourself.

For more information call **800-223-5688**. In New York call **(212) 751-2300**. Or see your travel agent or write Dept. SA-1.

*Registered in Sweden

TO LEARN IS TO LIVE.

SALÉN LINDBLAD CRUISING
133 East 55th Street New York, N.Y. 10022

ANTIQUITY

A Periodical Review of Archaeology
edited by Glyn Daniel

Founded in 1927 by O. G. S. Crawford, ANTIQUITY has appeared regularly ever since and won acclaim the world over as the most authoritative journal in its field. While written by specialists, the articles, notes and reviews are popular in character and indispensable to all interested in the development of man and his past.

Professor Glyn Daniel, Faculty of Archaeology and Anthropology in the University of Cambridge, and Fellow of St John's College, has been Editor since 1956.

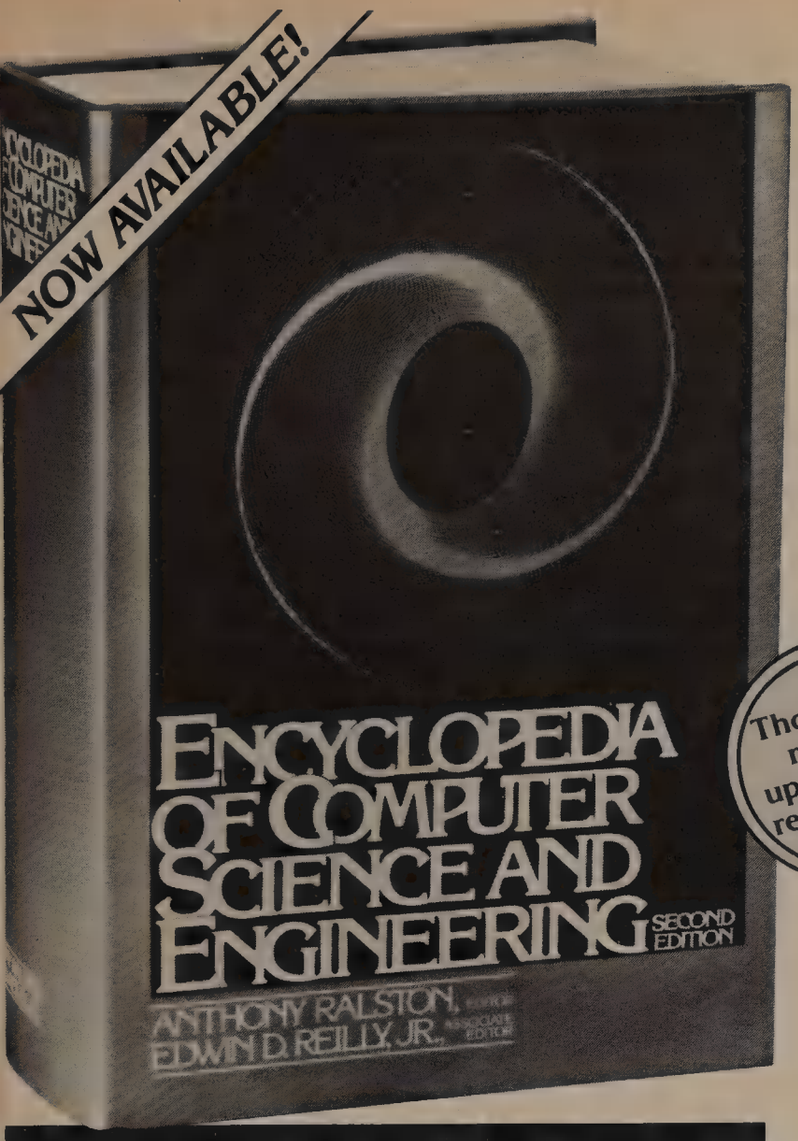
The annual subscription, postage included, is $35. Subscription forms and bankers' orders are available on request from

ANTIQUITY PUBLICATIONS LIMITED
Heffers Printers Ltd, King's Hedges Road,
Cambridge, England CB4 2PQ

The game of sonic disassembly and sequential, chemically controlled reassembly of these membranes and the structures they bear has been enthusiastically played for 20 years. There are still, however, missing pieces in the puzzle of the site and nature of ATP production. The tiny protein globules on short rooted stalks that line the inner membrane offer one example. It seems likely that the three parts of this structure—root, stalk and head—constitute the final synthesizing enzyme, although hard proof is lacking. If the first seven chapters outline the classical biochemistry and the later particulate study of the chemical function of the organelles, chapter 9 sounds a strong new theme: the nature of the membranes themselves. These are no inert walls. It is a surprise to see how easily a single active enzyme, the important cytochrome oxidase, is quite ready to self-assemble in vitro into little membrane bags on its own, given a good dollop of phospholipids. Structure and function flow as one within the mitochondria.

The two final chapters play very modern music indeed. The population of mitochondria in a cell divide and fuse to meet the cell's energy needs by processes little understood and only sketchily reviewed here. These extranuclear granules hold a DNA instructional legacy all their own. Here is a beautiful micrograph of the DNA circlet of a human mitochondrion, a simple tape loop a few microns long, compared with the meter or two of tape intricately wound and unwound within each cell nucleus. That singular piece of the human genetic fabric is now being mapped both geneticallly and for its nucleotide sequences. Yeast and mammalian mitochondria differ considerably in detail, yet both are remarkable in that they rely on an unexpected three-nucleotide coding symbol for translation into one of the essential amino acids. That essential auxiliary tape of every human cell does not share the substitution code for protein synthesis otherwise universal in life.

These final chapters are rich in the results and methods of today's molecular biology of information; the book carries the story up to late 1980. The text ends with a too terse account of the conjectural origins of the mitochondria, which give many signs of alien origin, quasi-bacterial live-in servants within the larger cell. The experienced author is not inclined to speculation, but mitochondrial studies clearly tax the triumphant techniques of molecular biology today. It is satisfying if hard going to follow these investigators well beyond the easier generalizations in the introductory textbooks. The text is attractive and clear; somehow, probably by way of some mutant editing disk, "energy conversion" has throughout received the reading "energy conservation."



TAKE THE *NEW* ENCYCLOPEDIA OF COMPUTER SCIENCE AND ENGINEERING (AN $87.50 VALUE) FOR $2.95!

Thoroughly revised, updated and reorganized.

When you join the Library of Computer and Information Sciences. You simply agree to buy 3 more books—at handsome discounts—within the next 12 months.

Just published, the ENCYCLOPEDIA OF COMPUTER SCIENCE AND ENGINEERING is the most complete computer information resource available. Indispensable to consultants, business people, data processing professionals, and enthusiasts, it is a veritable data base of information on:

- Hardware
- Software
- Programming languages
- Artificial intelligence
- Operating systems
- Computer applications
- Personal computing
- and much more!

The Encyclopedia is organized to make finding and using its wealth of information an ease. Articles are alphabetically arranged and are cross-referenced to related articles and to specific subject matter. The clear and expanded appendices include abbreviations, acronyms, special notation and terminology, as well as numerical tables, the mainstay of applied technologies. A complete 5,000-term index contains references to sub-categories, doubles as a computer science dictionary, and is an invaluable tool for locating specific information.

Praise For the First Edition:
Called "Impressive . . . comprehensive . . . well done" by *Datamation*, and ". . . a real treasure cache" by *Business Management*, the new Second Edition promises to eclipse *Computer Management's* statement on the original of "There isn't another book like it."

Send for your *free* 10-day trial look today.

The Library of Computer and Information Sciences is the oldest and largest book club especially designed for the computer professional. In the incredibly fast-moving world of data processing, where up-to-date knowledge is essential, we make it easy for you to keep totally informed on all areas of the information sciences. In addition, books are offered at discounts up to 30% off publishers' prices.

Begin enjoying the club's benefits today!

This up-to-date Second Edition contains:

- **Over 1,670 pages of revised, expanded, and updated text**
- **550 articles on virtually every aspect of the computer sciences**
- **Written by 301 distinguished authorities (over 90 experts have been added to the list of ENCYCLOPEDIA contributors)**
- **Profusely illustrated with over 500 photos, and over 250 diagrams, graphs and charts**

4 Good Reasons to Join

1. The Finest Books. Of the hundreds of books submitted to us each year, only the very finest are selected and offered. Moreover, our books are always of equal quality to publishers' editions, *never* economy editions.
2. Big Savings. In addition to getting the ENCYCLOPEDIA OF COMPUTER SCIENCE AND ENGINEERING for $2.95 when you join, you keep saving substantially—up to 30% and occasionally even more. (For example, your total savings as a trial member— including this introductory offer—can easily be over 50%. That's like getting every other book free!).
3. Bonus Books. Also, you will immediately become eligible to participate in our Bonus Book Plan, with savings up to 70% off the publishers' prices.
4. Convenient Service. At 3-4 week intervals (16 times per year) you will receive the Book Club News, describing the Main Selection and Alternate Selections, together with a dated reply card. If you want the Main Selection do nothing and it will be sent to you automatically. If you prefer another selection, or no book at all, simply indicate your choice on the card, and return it by the date specified. You will have at least 10 days to decide. If, because of late mail delivery of the News, you should receive a book you do not want, we guarantee return postage.

If the reply card has been removed, please write to
The Library of Computer and Information Sciences
Dept. 7-BN8, Riverside, N.J. 08075
to obtain membership information and an application.

Established 1845 SCIENTIFIC AMERICAN January 1983 Volume 248 Number 1

The Large-Scale Cultivation of Mammalian Cells

Novel reactors have been designed for growing in culture large quantities of the fragile, complex cells that synthesize medically important proteins such as interferon and monoclonal antibodies

by Joseph Feder and William R. Tolbert

There are molecules of great investigative, clinical and perhaps commercial value that can best be produced by growing in culture the human cells and other mammalian cells that synthesize them. The problem is that it is no simple matter to grow large quantities of mammalian cells in an artificial medium. The well-developed technology of industrial microbiology is adapted to the requirements of bacteria, yeasts and molds. Each single cell, encased in a tough cell wall, is an independent metabolic factory with fairly simple nutritional requirements; for bacteria, glucose and some simple salts will often suffice. Microorganisms grow well floating free in a liquid culture medium in tanks with a capacity of as much as 50,000 gallons, resisting damage even when they have proliferated to form a thick suspension and even when the suspension is stirred vigorously with a mechanical agitator.

Mammalian cells are different. They are larger than most microorganisms, more fragile and more complex. The delicate plasma membrane that encloses an animal cell is not encased in a tough cell wall. The mammalian cell's nutritional requirements are more stringent than those of most microorganisms and indeed have not yet been fully defined. Rather than being a free-living organism, a mammalian cell is adapted to a specialized life as part of an organized tissue, dependent on the specialized functions of many other cells and on a circulatory system that ensures a precisely adjusted and stable environment for each cell. Such a cell resists being separated from its tissue and grown in an artificial medium. Most animal cells will not grow at all in suspension; they grow only when they can attach themselves to a surface. Over the years techniques have been developed for growing mammalian cells on a small scale in the laboratory. It has proved to be much more difficult to grow them efficiently on even a moderately larger scale.

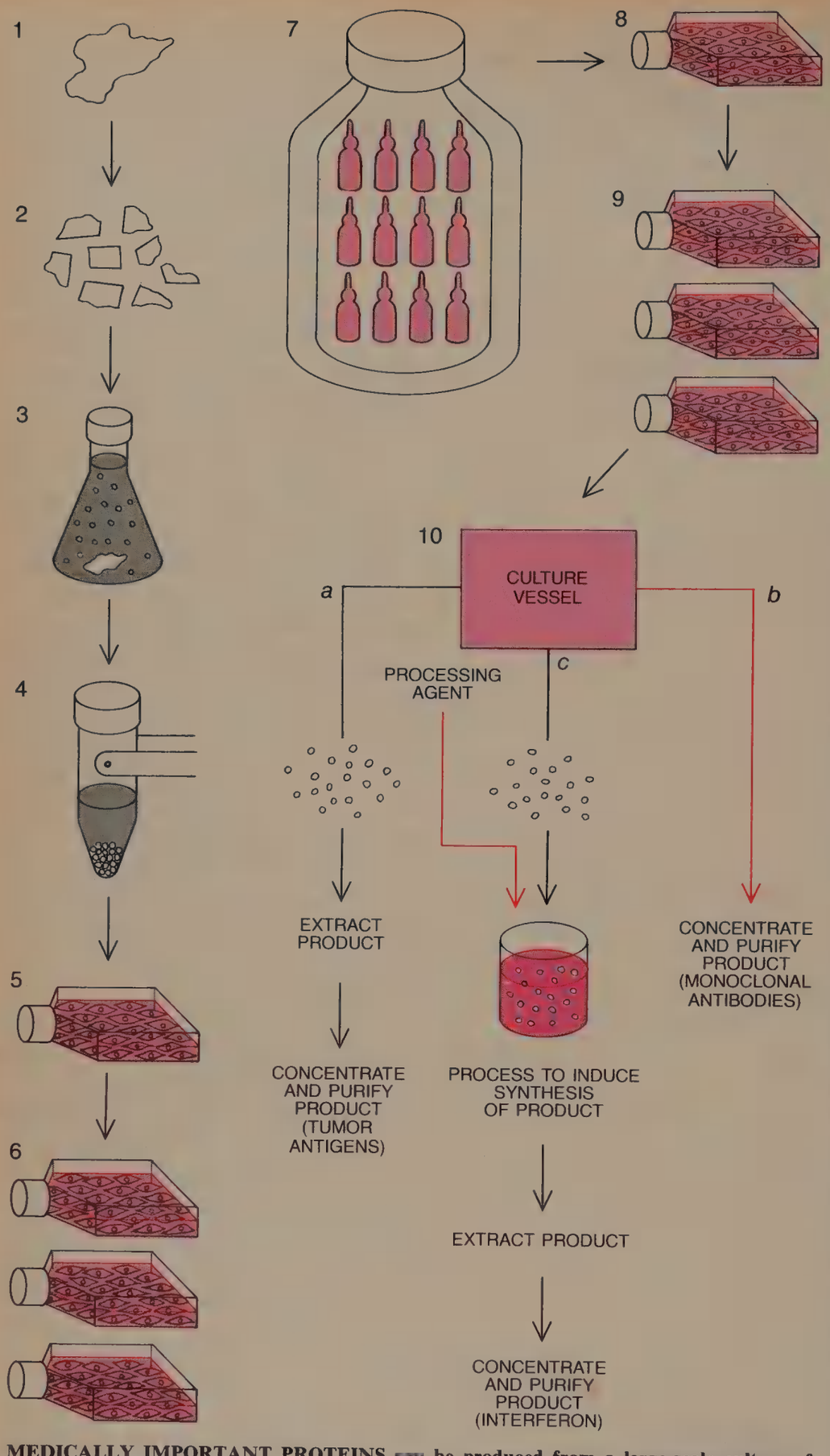
There is a pressing need for large-scale culture methods. Consider interferon. It is released by animal cells and inhibits viral infection. It was discovered in 1957, but its clinical efficacy is still not well established, mainly because it has been hard to grow large quantities of the mammalian cells that can be induced to synthesize it. By the same token, monoclonal antibodies targeted against a specific protein are manufactured by "hybridomas" formed by the fusion of an antibody-secreting cell to a malignant myeloma cell. Mouse hybridomas can be grown efficiently as tumors in laboratory animals. Human monoclonal antibodies promise to be powerful therapeutic agents, but the human hybridomas that make them need to be grown in culture.

Human urokinase, an enzyme that activates plasminogen to form the blood-clot-dissolving enzyme plasmin, has obvious clinical promise. A good source of urokinase is human kidney cells, which are difficult to grow on a large scale. Angiogenic factors are cell mediators that promote the growth of blood vessels. Specific tumor angiogenic factors released from tumor cells appear to promote the vascularization and growth of solid tumors. Large amounts of these factors are needed for research that may lead to the development of drugs for regulating angiogenesis (and perhaps for controlling the growth of a cancer); large amounts may also be useful for direct administration to promote revascularization. Human cells carry cell-surface antigens: molecules that serve as markers distinguishing self from nonself and tumor cells from normal cells. Tumor antigens are the subject of much current research, and they can be made available in quantity only by large-scale cell culture.

The demand for these molecules and for many others has motivated recent efforts to develop efficient large-scale culture systems for mammalian cells. The efforts have not been discouraged by awareness of the dramatic advances in recombinant-DNA technology, which make it feasible to introduce the gene encoding a desired protein into bacteria and then induce the bacteria to synthesize the protein. Many biologically important molecules cannot be synthesized by these methods. And in the case of a substance, such as the angiogenic factor, that has not even been characterized, the only way to get enough material from which to isolate it is to grow the cells that make it.

A mammalian-cell culture begins with a mammalian tissue. The tissue is dissociated either mechanically or enzymatically, or by a combination of the two methods, to yield a mixture of single cells and small clumps of cells. The mixture is inoculated into an appropriate liquid growth medium that ordinarily includes salts, glucose, certain amino ac-

CULTURE VESSELS for growing mammalian cells in quantity are shown in the authors' laboratory at the Monsanto Company. The three cylindrical steel tanks are 100-liter reactors for growing cells that do not need to be anchored to a surface and can be grown in suspension in a culture medium. The glass vessels on the shelves are spinner bottles in which cells are grown before being inoculated into a large-scale culture vessel. Each bottle stands on a magnetic stirrer. The red color of the culture mediums results from an acid-base indicator.



MEDICALLY IMPORTANT PROTEINS can be produced from a large-scale culture of mammalian cells by the general procedure diagrammed here. A tissue (*1*) is cut into small bits (*2*), which are treated with a protein-dissolving enzyme (*3*) that releases individual cells. The cells are collected by centrifugation (*4*) and inoculated into a nutrient medium, where they proliferate until they cover the floor of the culture flask (*5*). The confluent cells are released from the floor of the vessel by an enzyme and are reinoculated into several flasks to expand the culture (*6*). A stock of the cells is stored frozen in liquid nitrogen (*7*). As required, some of the cells are thawed to start a new culture (*8*), which is expanded (*9*) to make enough cells to inoculate a large-scale culture vessel (*10*). The desired proteins can be retrieved in several different ways. The cells can be harvested and a cell product such as the surface antigens of tumor cells can be removed and purified (*a*). If the desired product is secreted by the cells, as it is in the case of monoclonal antibodies, it can be removed from the medium, which is pumped out of the reactor while the cells are retained (*b*). In some instances an extra stage is required in which the cultured cells are treated to induce the synthesis of a desired product, such as interferon (*c*).

ids (the subunits of proteins) and blood serum (usually accounting for between 5 and 20 percent of the medium).

The serum is included as a means of providing components that have not yet been identified but have been shown to be needed if the cells are to live and grow in an artificial medium. Serum is expensive, and its cost largely determines the economic feasibility of a particular culture system. Cells in culture are deprived of the elaborate immune defense system that is an integral part of the intact animal, and so antibiotics are often included to prevent infection of the culture. The *p*H, the temperature and the level of oxygen and carbon dioxide must be closely controlled. The salt content is regulated to keep the liquid medium's osmotic pressure within bounds and preserve the integrity of the fragile cell membrane.

Although most mammalian cells need to be anchored to a solid surface, cells that originate in blood or lymphatic tissue, along with most tumor cells and other transformed cells, can be adapted to growth in suspension. Such cells can be cultured in spinner bottles. These are simple glass containers in which a stirring element is suspended; the element is driven by an external rotating magnet at the base of the flask. Spinner bottles with capacities of from 25 milliliters to 15 liters are available for laboratory applications.

For anchorage-dependent cells (most normal cells) a number of laboratory vessels are appropriate, ranging from "microwell" plates and flat-bottomed Petri dishes through tissue flasks of various sizes. The most effective vessel is the roller bottle: a cylindrical vessel that is oriented with its long axis horizontal and is continuously rotated. Cells attach themselves to the inner surface of the cylinder, and the slow rotation (about one revolution per minute) exposes them alternately to the liquid medium and to air. The largest roller bottles offer a surface area of some 1,600 square centimeters.

For cells that grow in suspension large-scale reactors have been developed by modifying the fermenters in which single-celled microorganisms are grown. In the process, however, certain inefficiencies have been preserved. Most fermenter-style reactors are elaborate and expensive. For asepsis they rely on complex steam-sterilization equipment and the inclusion of antibiotics in the medium. The large-scale growth of anchorage-dependent cells presents even harder problems, which until recently have been avoided by simply multiplying the number of conventional vessels, most commonly roller bottles.

Some years ago our laboratory in the research department of the Monsanto Company undertook to devise new sys-

tems for large-scale cell culture that would meet our pressing need for a dependable source of hundreds of liters of cell suspensions each week. We decided at the outset that the systems must maintain absolute sterility without depending on antibiotics, which can mask the slow development of a low-grade infection in a culture for some time and can affect the metabolism of cultured cells in subtle ways that cannot always be predicted. We also decided on a modular approach: individual culture vessels would be limited in size and large-scale production would be attained simply by bringing more modules on line. Finally, we were determined to explore innovative designs that would address the specific growth requirements of the large, fragile mammalian cell.

A simple and inexpensive 100-liter reactor for the growth of cells in suspension met all these criteria. It was made from a 30-gallon stainless-steel drum fitted with a removable head through which all input and output attachments were made. The reactor is small enough to be sterilized in an autoclave; a double-septum port in the removable head makes it possible to take samples as often as needed without contaminating the culture. The necessary agitation is provided by a vibromixer (a device that was first demonstrated in 1965 by George E. Moore of the Roswell Park Memorial Institute). A disk with conical apertures is attached to a vertical shaft that oscillates up and down at high speed. The vibration of the disk and the passage of medium through the holes combine to agitate the medium largely in a vertical rather than a horizontal direction. An added advantage of the vibromixer is that its motion is transmitted through a flexible diaphragm, providing an absolute, static seal at the top of the vessel. Tubes introduced through the top of the vessel serve to bubble oxygen through the medium and to move air enriched with carbon dioxide over the surface of the medium.

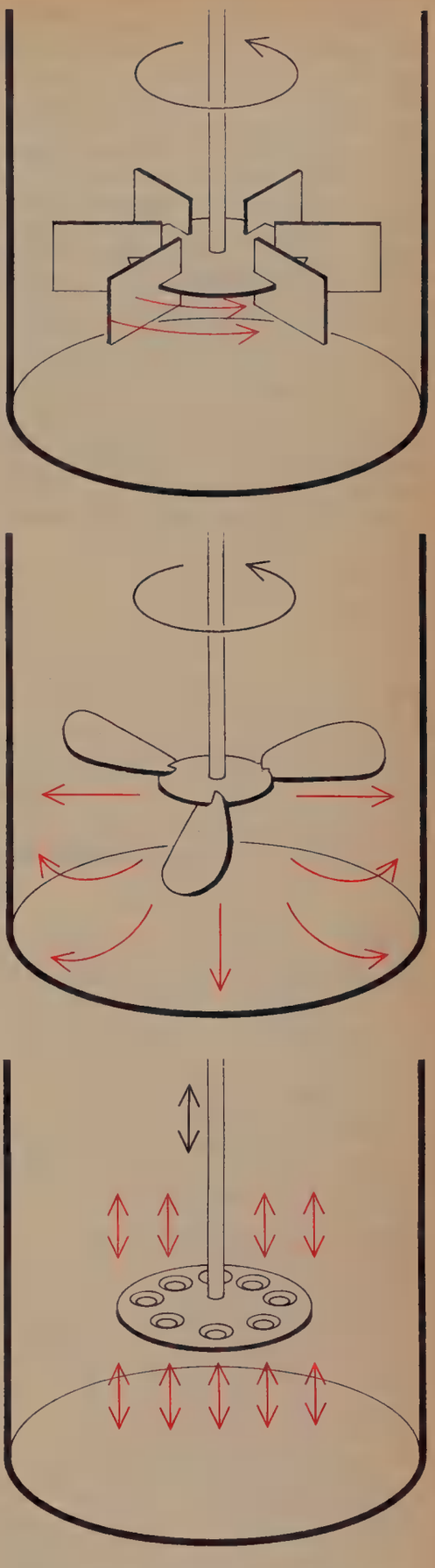
We built a number of these 100-liter vessels at about a twentieth of the cost of conventional fermenters. They are adaptable to a variety of cells that grow in suspension, and over the past few years hundreds of thousands of liters of human-cell and rodent-cell suspensions have been grown and harvested. Nevertheless, we decided against simply making larger versions of the 100-liter reactor in order to attain higher levels of production. Instead we set about developing a system that would more closely imitate the conditions under which a mammalian cell grows in nature.

A conventional cell-culture system, including our 100-liter vessel, exposes growing cells to a constantly changing environment. Cells are inoculated into a medium that is rich in just the right nutrients and is devoid of metabolic waste products; as the cells grow, the concentration of nutrients decreases and waste products accumulate. This imposes a limit on the density of the culture. In the living animal, on the other hand, homeostasis is the rule. An efficient circulatory system delivers oxygen and nutrients and removes carbon dioxide and waste products, providing every cell with a stable environment even though the cell density is between 500 and 1,000 times as high in a living tissue as it is in a typical culture medium. It would be desirable to provide cells in culture with a near-steady-state environment. A way to do that is to constantly perfuse fresh medium through the culture and constantly remove depleted medium. The trick is to renew the medium without removing the cells. In 1969 Philip Himmelfarb and Philip S. Thayer of Arthur D. Little, Inc., developed a laboratory-scale perfusion system for animal-cell suspension cultures and reported attaining much higher cell densities than could be achieved with conventional suspension systems.

We have developed a moderate-scale perfusion system with four-liter and 44-liter vessels in which cells are grown in amounts that compare with what is ordinarily grown in vessels having capacities of 100 and 1,000 liters. The perfusion is made possible by the introduction into the system of a satellite vessel containing a cylindrical porcelain filter with pores less than two micrometers (thousandths of a millimeter) in diameter. The filter retains the cells, which are pumped back into the culture vessel; some of the medium is recycled and some of it is withdrawn as fresh medium is fed into the system. To minimize clogging of the filter pores by cells or cellular debris, shear forces are established at the surface of the filter either by rotating the filter or by surrounding it with a rotating agitator. Special sensors, sampling ports and input and ouput piping make it possible to maintain the *p*H and the concentration of oxygen, carbon dioxide and nutrients at the proper level even when cells have proliferated to a very high density.

The high cell density achieved with perfusion made it necessary to design a new and gentler agitating mechanism. Four flexible sheets woven of a plastic monofilament rotate slowly (at from 8 to 20 r.p.m.) in the culture vessel, billowing like sails as the liquid medium spills from their trailing edge. By distributing the required energy of agitation over a large area the flexible-sail mechanism reduces the shear forces acting on individual cells.

To evaluate the efficiency of the perfusion system we compared the growth of rat tumor cells in it and in a conventional system and monitored the concentration of glucose and of lactic acid



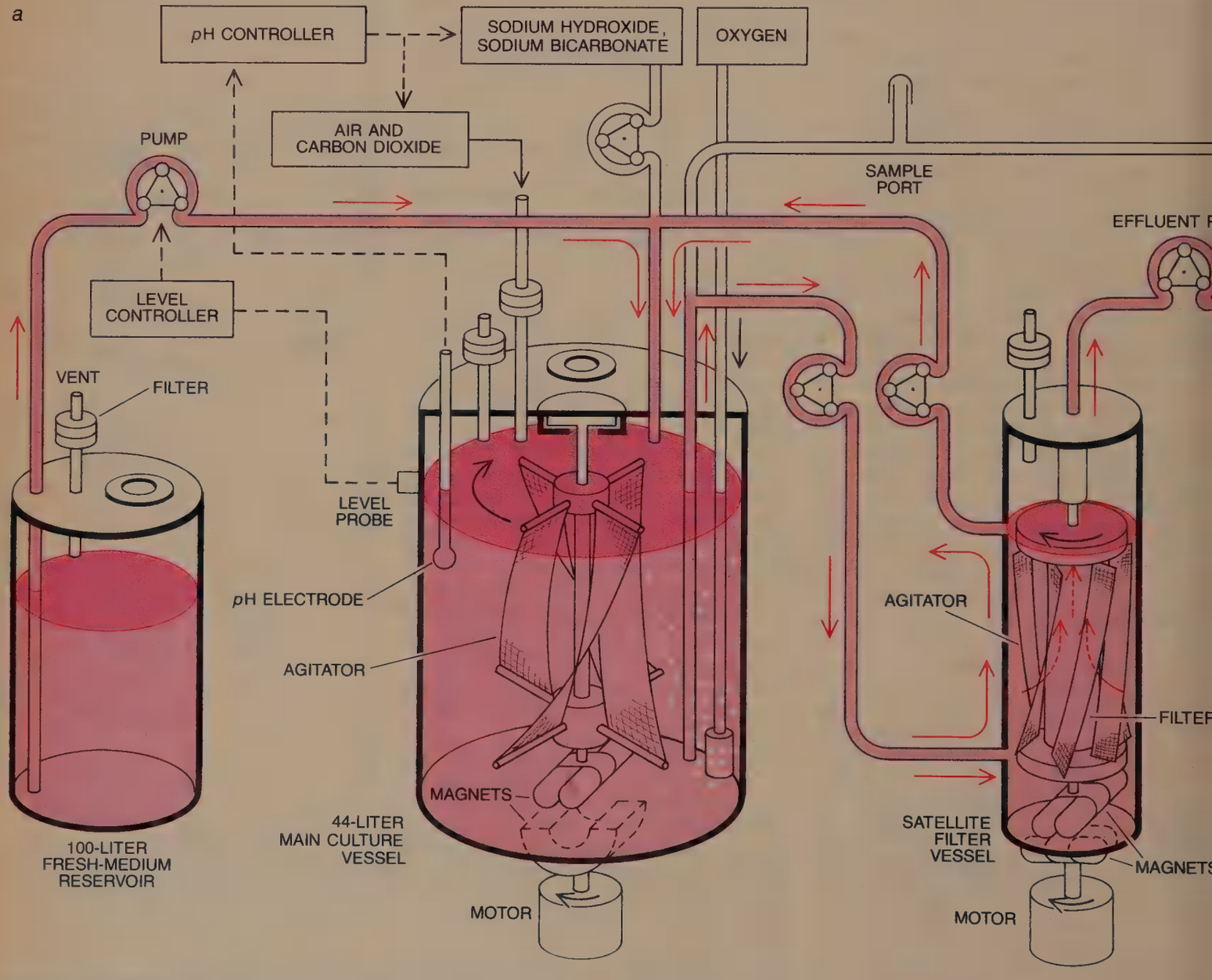
AGITATION is required to keep cells uniformly suspended in the medium and to distribute dissolved gases evenly. The turbine agitator (*top*) is standard in the fermenters in which microorganisms are grown, but its rapid rotation sets up strong horizontal shear forces that can damage fragile mammalian cells. An agitator shaped like a marine propeller (*middle*) provides both horizontal and vertical mixing at a lower rate of rotation. The vibromixer (*bottom*) oscillates vertically, increasing vertical mixing and thereby improving aeration of the medium and reducing cell damage.

(a major metabolic product) in the effluent medium. The rate of growth of the cells was about the same in the two systems but the cell density was much higher in the perfusion reactor: about 25 million cells per milliliter compared with a million cells per milliliter in the conventional reactor. At maximum density all the perfusion-system cells were still alive, whereas some 30 percent of the cells in the conventional one were dead. Perhaps most significant economically, the yield of cells per liter of medium expended was more than two and a half times as high in the perfusion system (even though the medium is constantly renewed in such a system). Moreover, we found that the perfusion reactor can be run in what is called the chemostat mode, in which some cells are continually pumped out of the system along with the expended medium, thereby maintaining a constant cell density at an optimum level.

The vast majority of normal mammalian cells, as we mentioned above, grow only when they are attached to the surface of a culture vessel. Roller bottles and tissue flasks cannot provide a large enough ratio of surface area to volume for practical large-scale culture. The area-to-volume ratio can be increased by means of various stratagems. Cells can be grown, for example, on spongy polymers, on arrays of thin tubing or hollow fibers, on stacks of thin plates or on microscopically small beads called microcarriers. For large-scale culture we have investigated two of these approaches: the hollow fiber and the microcarrier.

Studies done in our laboratory and by Richard A. Knazek and his colleagues at the National Cancer Institute showed that mammalian cells grow well on a variety of hollow fibers with an external diameter ranging from one-third to three-quarters of a millimeter. The fibers are porous, so that air passing through them diffuses to the cells on the outer surface, which is bathed in the culture medium. After studying a number of configurations we finally decided on a "flat bed" arrangement of the fibers. The core of the system is a very shallow bed of fibers, only three to six layers deep, to which the cells attach themselves and where they proliferate.

Fresh medium enters the stainless-

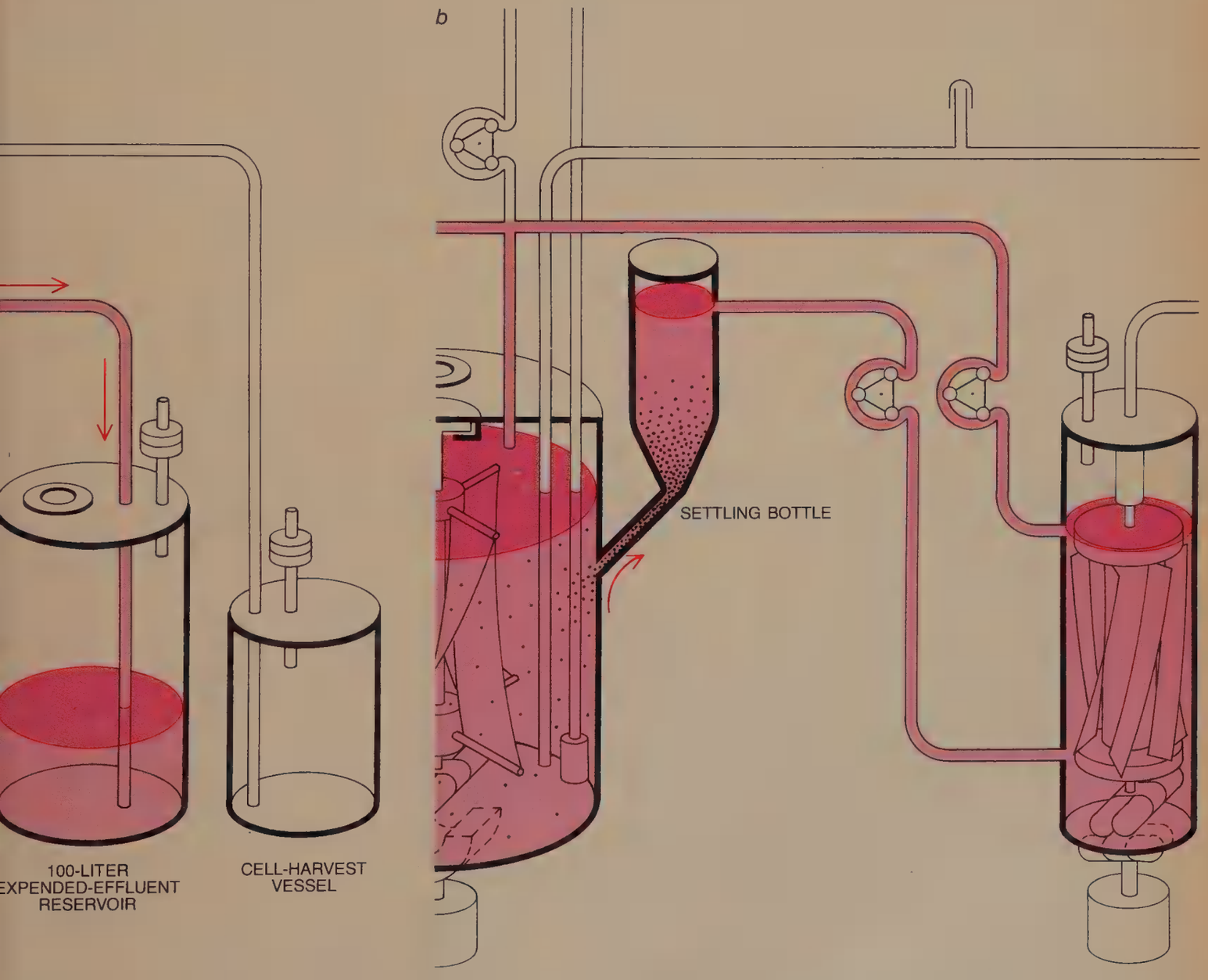


PERFUSION REACTOR designed by the authors moves fresh medium into the system and removes medium in which nutrients are depleted and waste products have accumulated, meanwhile retaining the cells within the system. The diagrams show the version of the reactor developed for cells that grow in suspension (*a*) and a modification (*b*) introduced for cells that grow on microscopic beads called microcarriers. In the main culture vessel gentle agitation is provided by four slowly rotating flexible sheets of monofilament nylon. Medium in the main vessel is continuously pumped to the satellite filter vessel, where some of the medium is pumped through the porcelain filter to the effluent vessel. The filter retains the cells, which are pumped, along with the remainder of the medium, back into the

steel housing below the fiber bed, perfuses up through the bed and is withdrawn at the top of the housing. To make the flow uniform and perpendicular to the plane of the fiber bed we pump the medium up to the bed through a micropore filter: a sheet of stainless steel with pores two micrometers in diameter. Another filter, with 20-micrometer pores, is placed above the bed to limit the amount of backflow. The advantage of the shallow-bed configuration is that the path of perfusion is very short (compared, for example, with the path in a cylindrical "cartridge" system, where the flow of medium parallels the axis of the fibers). As a result there is hardly any gradient from a high to a low nutrient level and from a low to a high waste-product level; all the cells are exposed to essentially fresh medium. The ends of the fibers are bonded together in a nontoxic, elastic material, which is then sliced in such a way that the hollow bore of each fiber is open to the chambers through which air and carbon dioxide are circulated.

We have built two prototype reactors, with fiber-surface areas of 930 square centimeters and 9,300 square centimeters. (The larger prototype is only 40 centimeters square and 4.5 centimeters high.) They have maintained high-density cultures for long periods of time. For example, the smaller prototype grew human lung cells to a density of about a million cells per square centimeter, or about 10 times the density that can be achieved in roller bottles. Perfusion of the medium is particularly effective in these reactors. It can be regulated to maintain a steady-state level of nutrients and wastes as the cells proliferate. That is done by monitoring the level of lactic acid in the expended medium, deriving from it an estimate of the number of cells on the fibers and adjusting the rate of flow of fresh medium accordingly.

The system does not just grow large numbers of cells; in effect it builds an artificial tissue as billions of cells proliferate to encrust the fibers and even penetrate them. Large quantities of cells are thereby maintained under conditions of minimal stress for a long time (from 21 to 59 days in our experiments), during which important molecules such as urokinase and angiogenic factor are produced continuously and are harvested



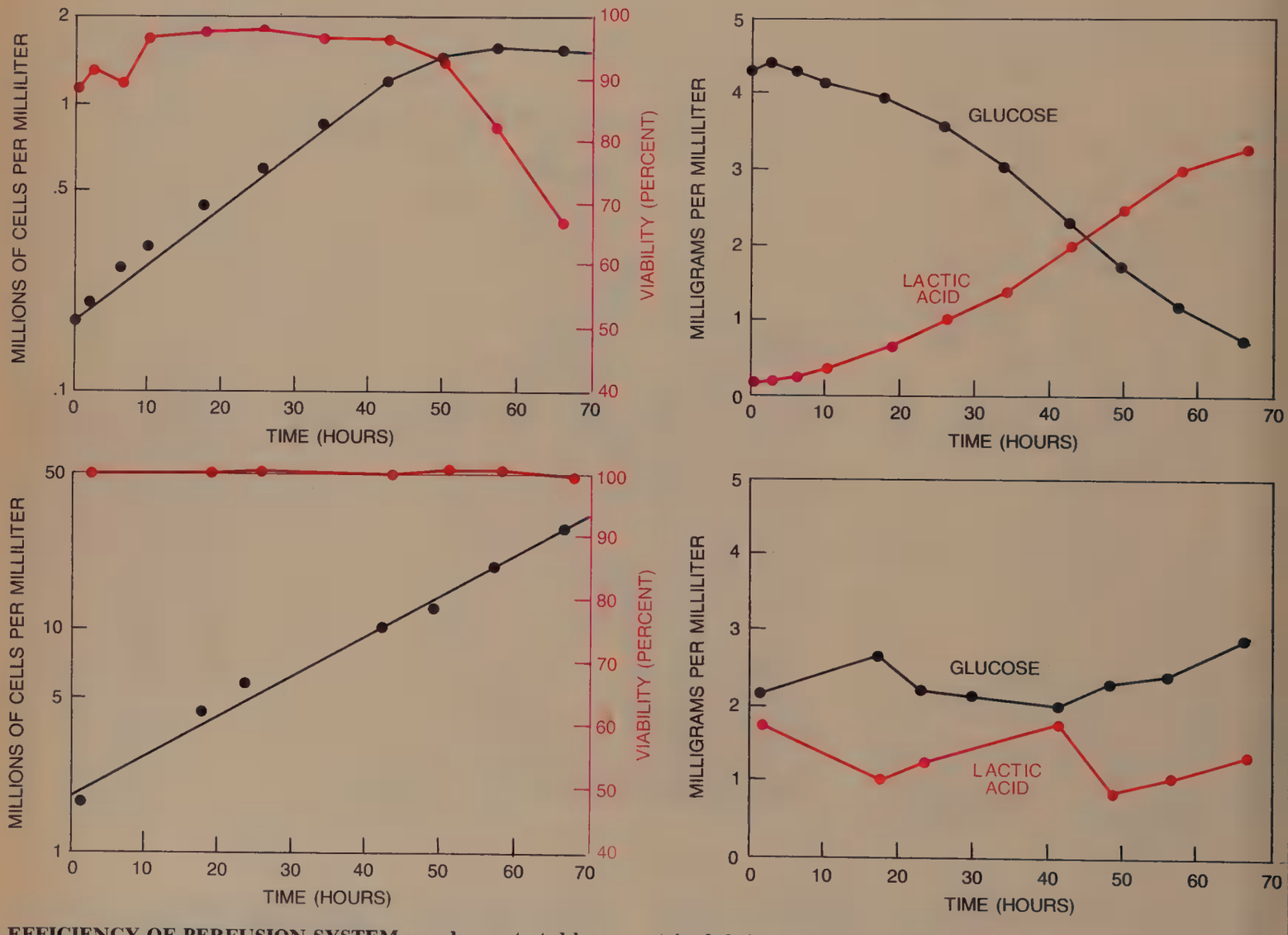
main vessel for further proliferation. A small version of the flexible-sail agitator rotates in the satellite vessel to minimize clogging of the filter pores. Substances secreted by the growing cells can be harvested from the medium in the effluent vessel. The cells can be harvested by being pumped out of the reactor into the cell-harvest vessel. Most normal mammalian cells grow only when they are attached to a surface rather than when they are in suspension in a liquid medium; microcarriers provide the necessary surface. To adapt the perfusion reactor for cells growing on microcarriers a settling bottle was placed in the line between the main vessel and the satellite vessel. Beads encrusted with growing cells settle out of suspension, forming a slurry that slides back into main vessel, where cells continue to proliferate.

from the spent medium. Careful control of contamination should make it possible to operate one of these systems almost indefinitely. Once maximum growth has been attained it is often feasible to substitute a simple maintenance medium, containing little or no serum, for the more expensive medium that supports cell proliferation, and to keep the cells at work producing a desired molecule.

The microcarrier method was first demonstrated in 1967 by Anton L. van Wezel of the Dutch National Institute for Public Health, and many variations of the approach have since been investigated and exploited for industrial production on a moderate scale. The microcarrier beads are composed of the natural glucose polymer dextran or one of various synthetic polymers and range from 50 micrometers to several hundred micrometers in diameter. They are suspended in a nutrient medium. Anchorage-dependent cells inoculated into the medium bind to the beads and multiply. The advantage of a microcarrier system is that although the cells are growing on a surface, the beads that provide the surface are suspended in the medium, and so the cells can be grown in conventional suspension vessels.

Such conventional systems are not, however, very efficient for microcarrier cultures. In addition to such usual concerns as controlling the concentration of nutrients and waste products, one must consider the amount and the distribution in the system of the microcarriers. Collisions between beads can injure cells. The collisions are increased by vigorous agitation, and so there has been a need for a gentler stirring mechanism to keep the beads in suspension without destroying too many cells. To grow large numbers of cells a large surface area is needed. That means a high density of beads, but too high a bead density makes for more collisions. The relatively high surface-to-volume ratio attained with microcarriers gives rise to yet another problem: as the cells proliferate there can be a very rapid depletion of nutrients and a parallel buildup of toxic waste products. Finally, there is the problem of scaling up from one growth vessel to a larger one. The encrusted cells must be removed from the beads, after which they are resuspended in the larger vessel with a new supply of beads. In conventional microcarrier systems it has been hard to remove the cells from the beads without damaging a large proportion of them.

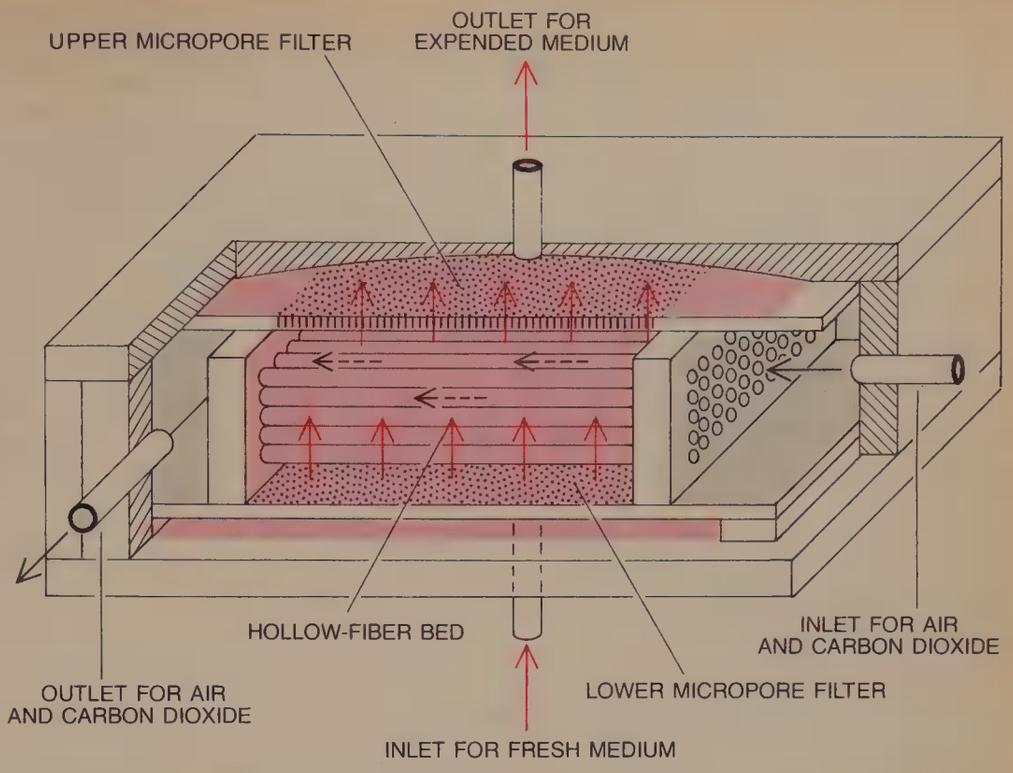
We adapted our novel perfusion system to handle microcarriers and found it made possible a significant increase in cell density. The flexible-sail agitators provide the necessary gentle stirring. Rotating at only from 8 to 20 r.p.m., they keep a high concentration of beads suspended in the medium (as much as 12 grams of beads per liter of medium) without causing cell damage. The major modification was the addition of a settling bottle between the main reactor vessel and the satellite vessel housing



EFFICIENCY OF PERFUSION SYSTEM was demonstrated by comparison of cell growth (*left*) and depletion of medium (*right*) in conventional spinner bottle (*top*) and in four-liter perfusion reactor (*bottom*). Rat tumor cells, which grow well in suspension, proliferated at about the same rate in both systems but attained much higher cell density in the perfusion system (*black curves at left*) and retained their viability even at high density (*colored curves at left*). The glucose and lactic acid levels reflect the depletion of nutrients and the buildup of waste products in the medium during cell growth. The glucose level fell steadily and the lactic acid level rose steadily in the conventional spinner bottle, whereas in the perfusion reactor virtually constant levels of nutrients and waste products were maintained.

the filter. The recycling pump moves the medium, with its suspended cell-encrusted beads, toward the satellite vessel. Once in the settling bottle, removed from the area where the medium is being stirred, the beads tend to settle and aggregate into a dense slurry that gradually slides back into the main vessel; medium that is essentially bead-free goes on to the satellite filter vessel and is either recycled or removed from the system.

In the settling process the beads and their cells come in close contact and continue to be perfused with fresh medium from the main vessel. Cells bridge from one microcarrier to another, forming large clumps of beads and cells, with the result that there can be up to four times as many cells as one would expect to find on a given bead-surface area. Another advantage, which we had not foreseen, arises from this highly aggregated condition: the separation of the cells from their microcarriers is facilitated, solving the scale-up problem. In the course of aggregation the attachment of individual cells to the beads is somehow reduced, to be replaced by multiple cell-to-cell attachments. When the clumped beads and cells are pumped out of the reactor, a brief treatment with the enzyme trypsin releases the cells from the microcarriers in good condition.

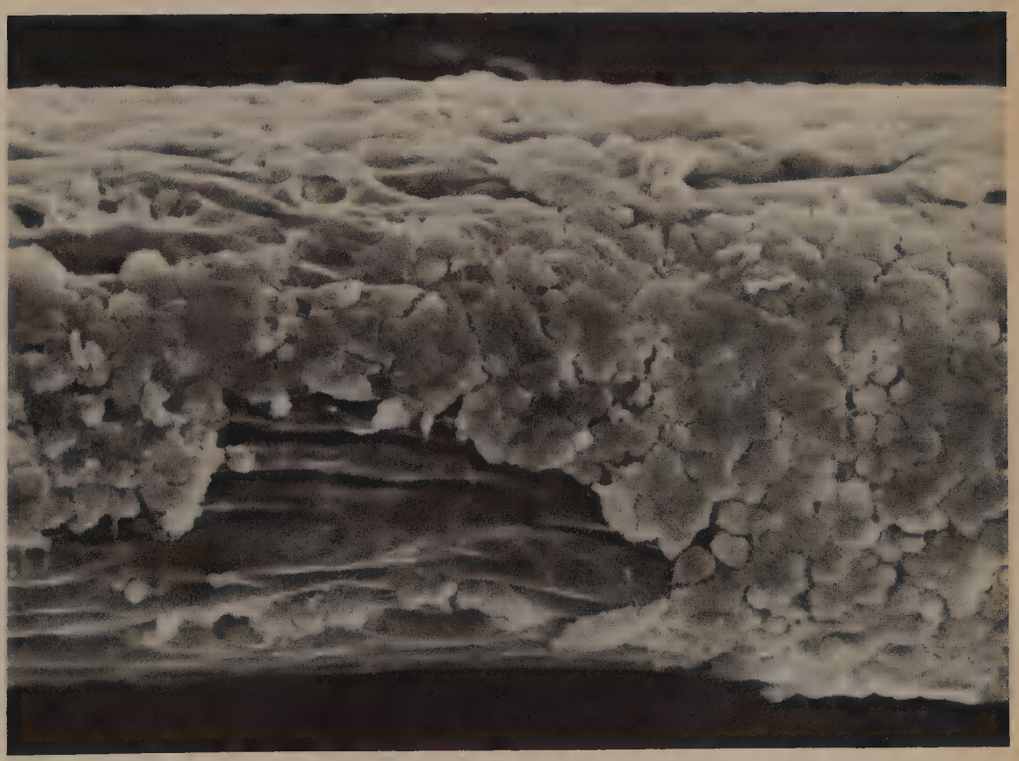


HOLLOW-FIBER PERFUSION REACTOR grows cells on a flat bed of porous plastic fibers .3 millimeter in diameter. The ends of the fibers are open to a supply of air and carbon dioxide, which moves through the fibers (*black arrows*); the outside of the fibers is bathed in a culture medium whose flow (*colored arrows*) is kept uniform by two micropore filters. Cells inoculated into the reactor attach themselves to the surface of the fibers and proliferate, in time forming what is in effect an artificial tissue that can be maintained for several months.

The microcarrier-perfusion reactor has proved to be an efficient system for the large-scale culture of anchorage-dependent cells. Microcarriers have always had the potential of providing a very large surface area. Now all that area can be exploited (actually overexploited, given the bridging effect) because perfusion maintains an optimum environment for cell growth even at extremely high cell densities, which in this system are from five to 10 times as high as before.

The combination of high-density cell growth and efficient scale-up leads to impressive results. In one experimental run of a small (four-liter) microcarrier-perfusion reactor we grew 40 billion human fibroblasts, the equivalent of a normal harvest from 1,300 roller bottles. Those cells were removed from their carriers and were inoculated in turn into a 44-liter reactor along with 400 grams of fresh microcarriers, which provided a surface area of 188 square meters. In the ensuing run 340 billion cells were grown, a harvest that would have required 11,000 roller bottles. The harvested fibroblasts served for the production not only of interferon but also of urokinase and an angiogenic factor. The best measure of the cost-effectiveness of the microcarrier-perfusion system is the yield of cells per liter of medium expended. In all our experiments the yield has been about four times as high as it is in roller-bottle systems.

Efficient large-scale systems for the culturing of a wide range of mammalian cells in the absence of antibiotics and at a reasonable cost are necessary to meet the needs of investigative laboratories as well as for production on a commercial scale. It seems clear that effective new technology can expand knowledge of biological processes and also provide significant quantities of cell products that promise to be important therapeutic agents.



HUMAN CANCER CELLS (HeLa cells) are seen growing on the surface of one of the hollow fibers in this scanning electron micrograph. In addition to eventually covering the entire surface of the fiber the cells penetrate its lobed structure. The enlargement is some 250 diameters.

Machines That Walk

Locomotion on legs resists imitation, but modern control technology should be able to solve the problem. Experiments with machines that hop and crawl can also illuminate the mechanisms of natural walking

by Marc H. Raibert and Ivan E. Sutherland

Many machines imitate nature; a familiar example is the imitation of a soaring bird by the airplane. One form of animal locomotion that has resisted imitation is walking. Can it be that modern computers and feedback control systems make it possible to build machines that walk? We have been exploring the question with computer models and with actual hardware.

So far we have built two machines. One has six legs and a human driver; its purpose is to explore the kind of locomotion displayed by insects, which does not demand attention to the problem of balance. The other machine has only one leg and moves by hopping; it serves to explore the problems of balance. We call the first kind of locomotion crawling to distinguish it from walking, which does require balance, and running, which involves periods of flight as well. Our work has helped us to understand how people and other animals crawl, walk and run.

Unlike a wheel, which changes its point of support continuously and gradually while bearing weight, a leg changes its point of support all at once and must be unloaded to do so. In order for a legged system to crawl, walk or run, each leg must go through periods when it carries load and keeps its foot fixed on the ground and other periods when it is unloaded and its foot is free to move. This type of cyclic alternation between a loaded phase, called stance, and an unloaded phase, called transfer, is found in every form of legged system. As anyone who has ridden a horse at a trot or a gallop knows, the alternation between stance and transfer can generate a pronounced up-and-down motion. We believe legged machines can be built that will minimize this motion.

Our work and related work by others may eventually lead to the development of machines that crawl, walk and run in terrain where softness or bumpiness makes wheeled and tracked vehicles ineffective and thus may lead to useful industrial, agricultural and military applications. The advantage of legged vehicles in difficult terrain is that they can choose footholds to improve traction, to minimize lurching and to step over obstacles. In principle the performance of legged vehicles can be to a great extent independent of the detailed roughness of the ground. Our objective has been to explore the computing tasks involved in controlling and coordinating leg motions. It is clear that very sophisticated computer-control programs will be an important component of machines that smoothly crawl, walk or run.

As we have indicated, locomotion is possible with or without dynamic balance. The animals that crawl avoid the need for balance by having at least six legs, of which at least three can always be deployed to provide a tripod for support. High-speed motion pictures of insects show that they commonly crawl with an alternating tripod gait.

Although a crawling machine that does not need dynamic balance can be built with four legs, such a machine performs awkwardly because its weight must be shifted at each step to keep it from tipping over. Satisfactory performance without active balance calls for at least six legs, since six is the smallest number of legs that always provide a tripod for support even when half of the legs are elevated. Several six-legged machines have now been built, each differing in size and in mechanical design. All of them depend on computer control of the legs.

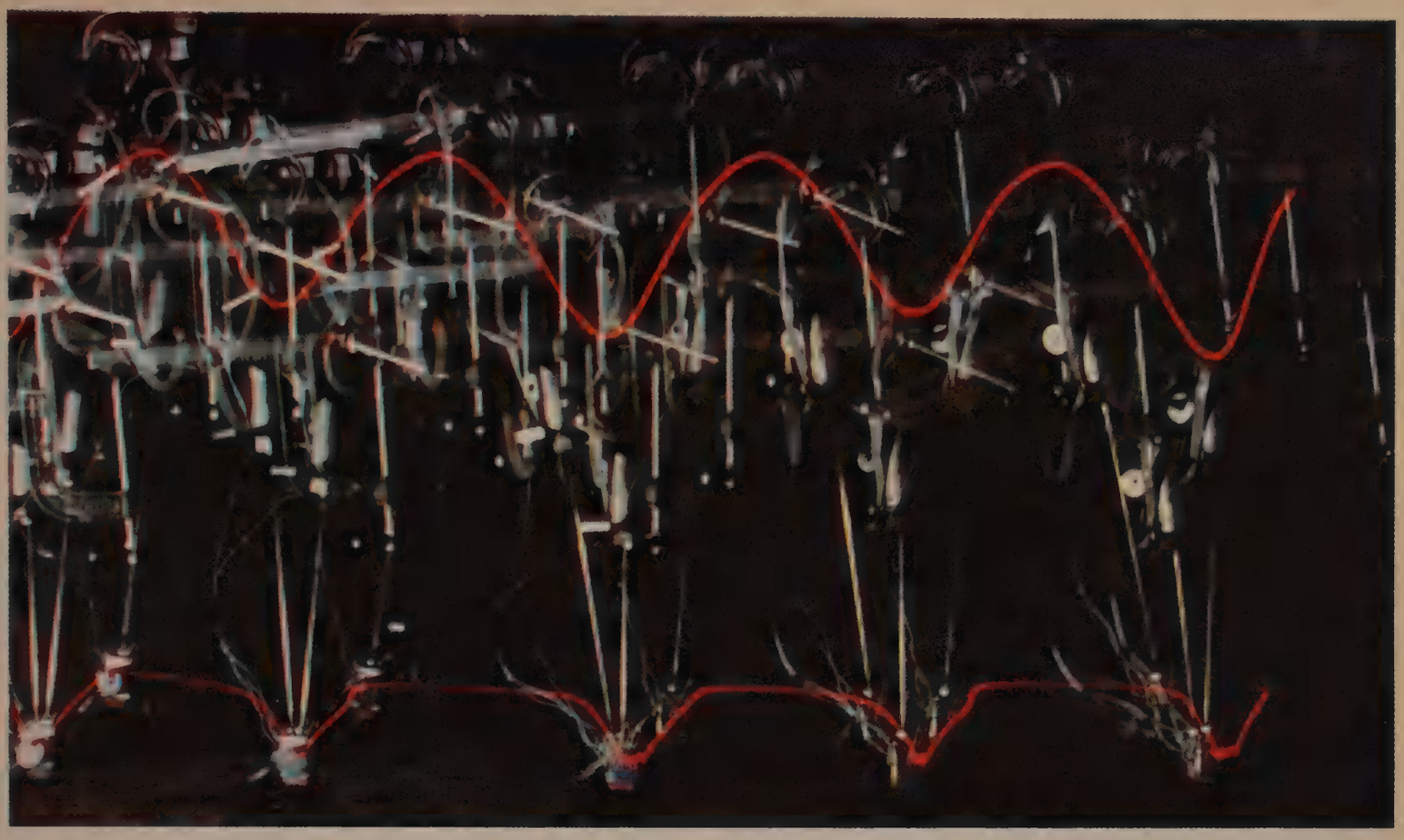
A computer program that controls such a machine accomplishes five tasks. First, it regulates the machine's gait, that is, the sequence and way in which the legs share the task of locomotion. Six-legged machines work with gaits that elevate a single leg at a time or two or three legs simultaneously.

The simplest gaits involve a regular sequence of leg motions. A gait can be described by noting the sequence. For example, the tripod gait can be recorded as (1,5,3;6,4,2;), with the commas designating the concurrent use of legs and the semicolons sequential use. Similarly, gaits that elevate a single leg at a time such as (3;2;1;4;5;6;) and (3;4;2;5;1;6;) are useful. A gait that elevates several legs at once generally makes it possible to travel faster but offers less stability than a gait that keeps more legs on the ground.

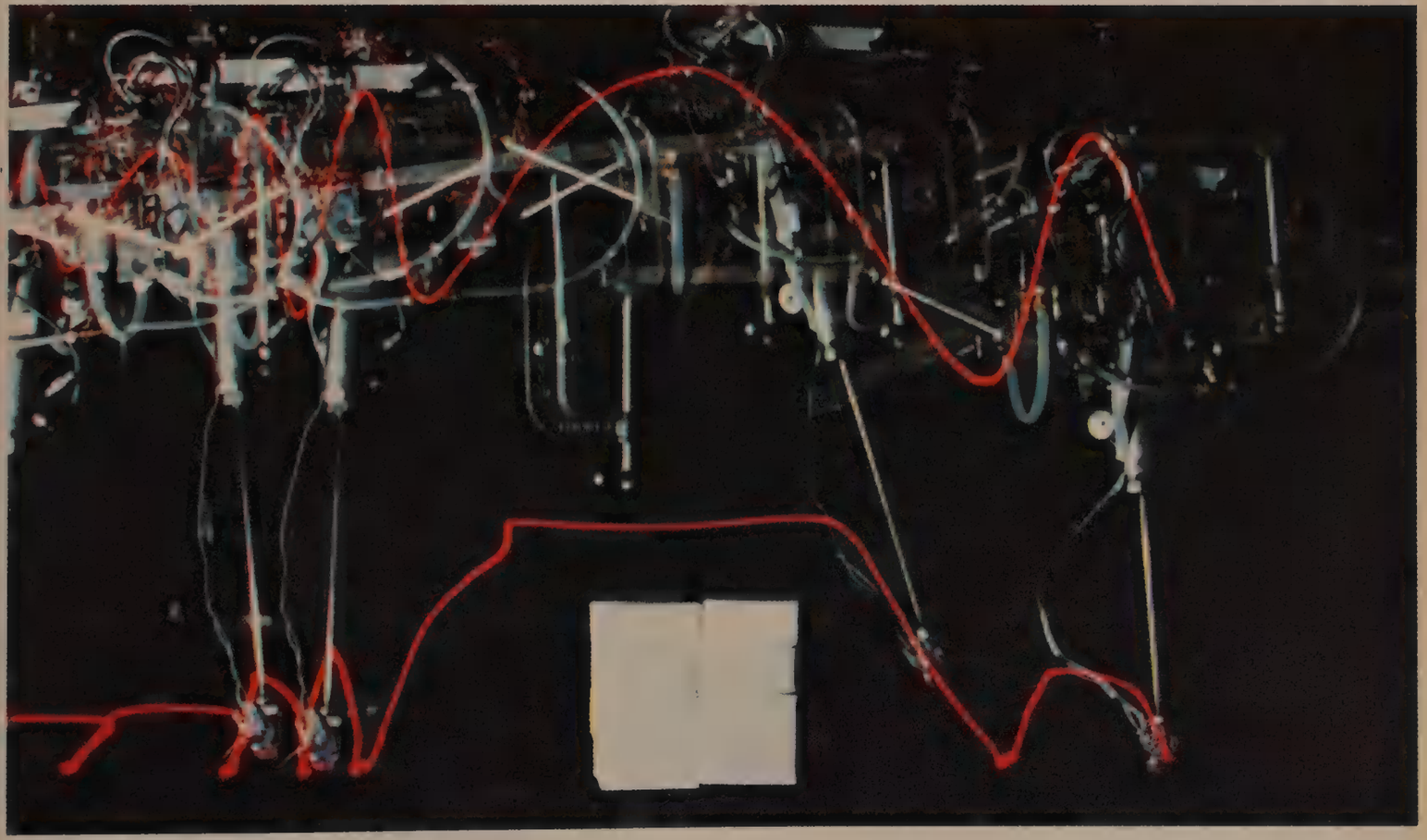
A second task of a computer program controlling a crawling machine is to keep the machine from tipping over. If the center of gravity of the machine moves beyond the base of support provided by the legs, the machine will tip. The computer must monitor the location of the center of gravity of the machine with respect to the placement of the feet to ensure that the base of support is always large enough. For simple gaits the geometry of the legs may suffice to keep an adequate base of support, but for more complex gaits a careful computation of static stability may be critically important.

Since many legs share the support of the machine, a third task of the control computer is to distribute the support load and the lateral forces among the legs. In the tripod gait, of course, the distribution of the support load is set by the geometry of the three supporting legs. With more than three supporting legs, however, the control computer must decide how to manage the distribution of loading in order to achieve higher-level objectives such as smoothness of ride and minimal disturbance of the ground.

Even when only three legs are supporting the machine, the control program must distribute the lateral foot forces. One way of looking at this task is to consider that the control system must keep the machine from simply doing isometric exercises against the ground. The amount of sensing and computation that is needed to distribute the lateral loads among many legs can be formidable. We have reduced this burden for the crawling machine we are building by providing passive hydraulic circuits that automatically distribute the sideways loads.



HOPPING MACHINE was built by one of the authors (Raibert) to explore the problems of controlling a machine that must balance as it moves. Its leg is actuated by compressed air and its motions are controlled by a computer that obtains feedback from position sensors. This version is held by a tethering arm and so balances in a single plane. A series of hops is recorded in this photograph, which was made while the camera lens was kept open. Red lights attached to the body and foot of the machine delineate the path of the hops.



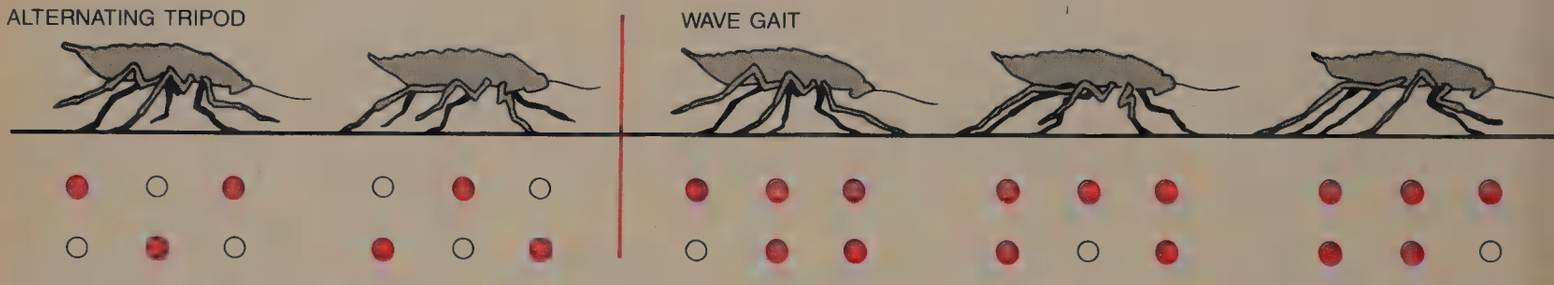
JUMPING MODE of the hopping machine shows it leaping over an obstacle. The machine approaches the obstacle from the right. When it is one step away, the operator pushes a "leap" button. As a result the maximum tension is generated in the drive actuator so that the altitude of the next hop will be increased. In flight the leg is shortened and its normal swinging motion is delayed to provide better clearance of the obstacle. A servomechanism controlling balance moves the leg to the correct landing angle and the leg is lengthened in preparation for landing. Thereafter the machine continues its normal hopping. The obstacle was 15 centimeters (six inches) high.

A fourth task of the control computer is to make sure the legs are not driven past the limits of their travel. The geometry of the legs may make it possible for one leg to bump into another; if legs can collide, the computer must limit their motion to prevent damage. To maximize the usefulness of each leg its placement on the ground must take into account the limits of the leg's motion and the expected motion of the machine during that leg's stance period. For example, if the machine is turning to the right, the forward legs should be placed farther to the right so that their sideways travel can be accommodated during the turn. For a vehicle with autonomous control the placement of the legs can be based on the planned future path of the vehicle. For a vehicle with a human driver the proper placement of each leg requires a prediction of the driver's commands for the next stance period.

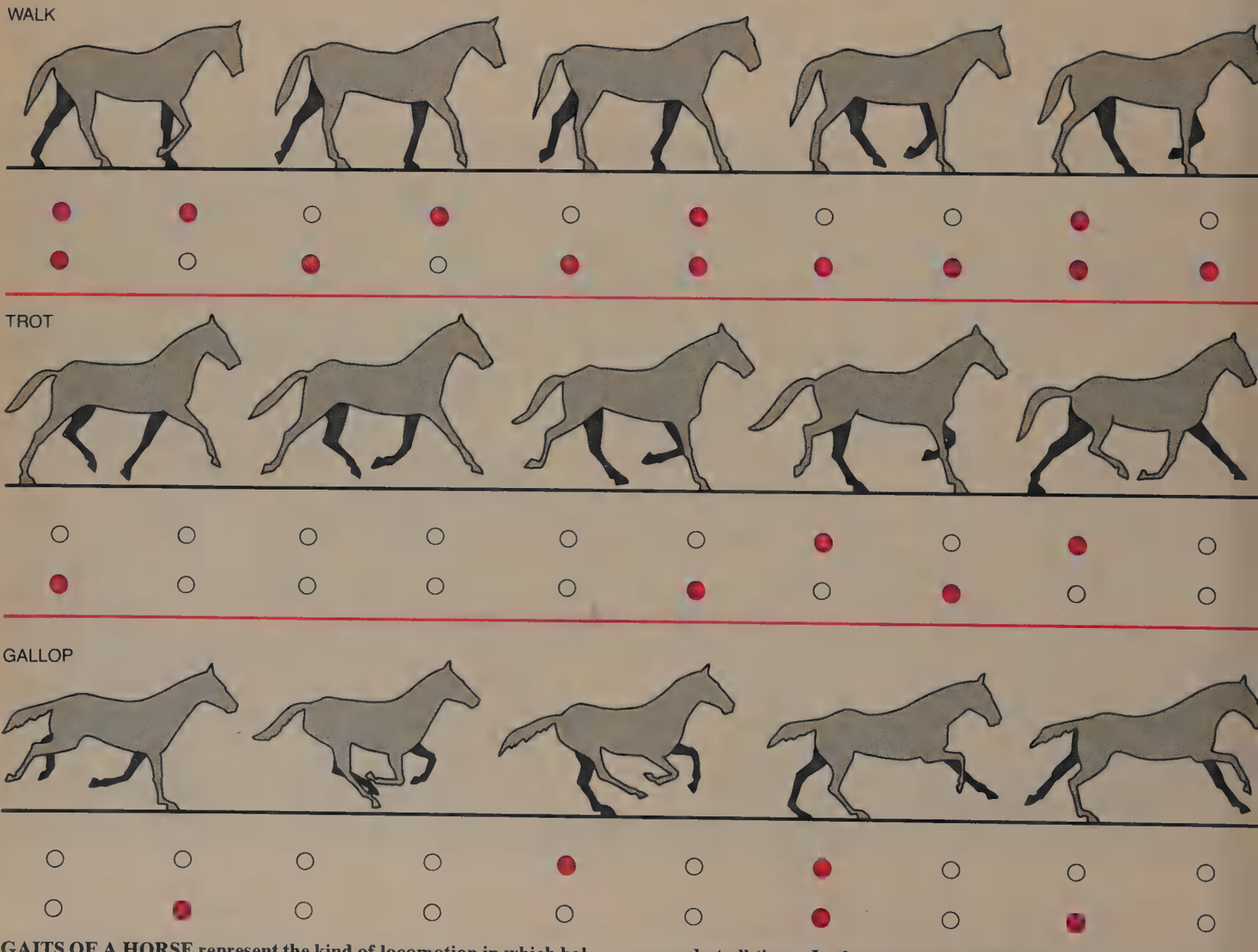
A fifth task for the control computer is to choose places for stepping that will give adequate support. On smooth ground the task is easy, but on rough terrain it may be exceedingly difficult. No system has yet been built that accomplishes this task. One can envision a terrain-scanning system that would survey the ground ahead of the machine and choose likely footholds. To make use of such a scanner the control computer would build an internal digital model of the terrain. Such a model would have to account only for bumps that are about the size of the machine's feet or bigger. Human input to the model might help in the evaluation of possible footholds.

One of us (Sutherland) is building a six-legged, hydraulically driven crawling machine. A gasoline engine provides its power and hydraulic actuators move its legs. There are six legs, so that dynamic balance is not needed.

A built-in microprocessor controls the legs by switching on or off the valves that regulate the flow of oil to the hy-



GAITS OF INSECTS provided a basis for designing the method of locomotion of the six-legged crawling machine portrayed on the cover of this issue. The circles below each drawing show whether the corresponding leg is on the ground or in the air: a filled circle represents



GAITS OF A HORSE represent the kind of locomotion in which balance is a factor. In the walk at least two of the horse's legs touch the ground at all times. In the trot and the gallop the animal periodically leaves the ground. The drawings are based on the stop-motion pho-

draulic actuators. Sensors in each leg report its position and the forces acting on it to the microprocessor. The machine is large enough to accommodate a human driver, who controls its speed and direction of motion and establishes the tilt of its body and its ground clearance. The vehicle's design speed is about two miles per hour.

One objective in the design of the vehicle was to minimize the amount of computation required to obtain a crawling movement. The hydraulic circuits are designed to make the legs move along useful paths without attention from the microprocessor, which merely selects one of the available paths for each leg. Thus the microprocessor is free to concentrate on selecting which legs to use for support and on deciding where to step next; it does not have to spend time computing the details of leg motion.

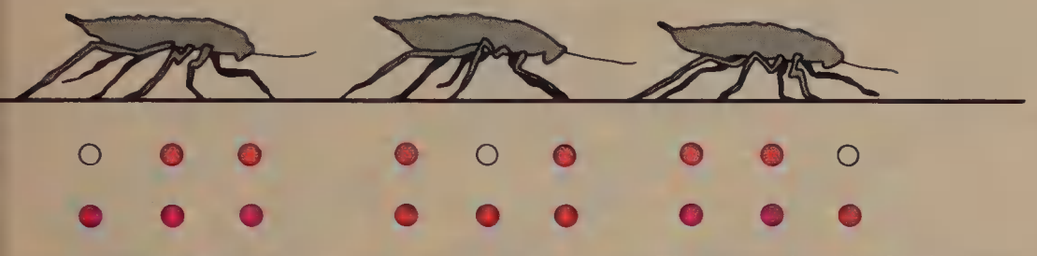
Each leg of the machine can swing fore or aft and up or down on the universal hip joint that attaches it to the frame of the machine. These motions are executed by lengthening or shortening the two hydraulic actuators per leg that are arranged in a *V* configuration above the leg. One setting of the valves provides that oil leaving one actuator will enter the other, so that as one actuator shortens, the other actuator lengthens by the same amount. Because of the geometry of the pivots this connection provides horizontal leg motion.

The horizontal motion can be powered or unpowered depending on the valve settings, so that some legs can serve to drive the machine forward while others coast. As legs are placed on the ground and accept load they are able to coast forward or backward as dictated by the motion of the legs already on the ground and driving. Hence the control computer does not need to compute the precise instant when a leg will touch the ground or the details of the motion required at the time of contact to obtain a smooth forward motion.

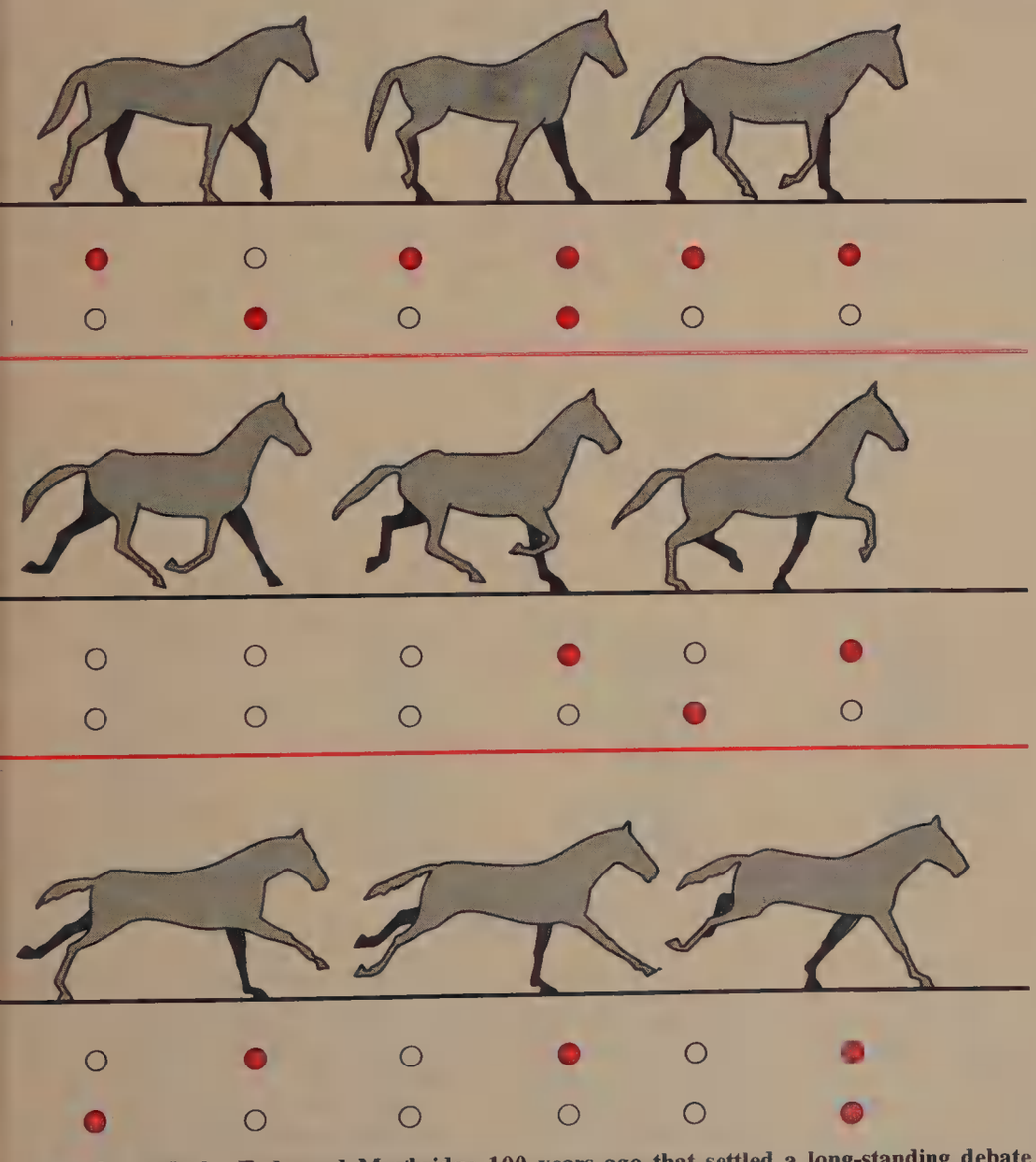
The knee joint of each leg is powered by a separate hydraulic actuator mounted horizontally along the leg. This actuator can be powered while the leg is raised in order to position the foot sideways for the next step. When the foot is on the ground, the knee joint must move slightly to match the circular path of the knee about the hip to the straight path of the foot on the ground. It is a complex motion, but it does not call for action by the computer; instead a simple parallel connection of the knee-joint actuators enables all the knees to accommodate to the average motion of the vehicle. An additional hydraulic pump provided in the system can force a collective sideways motion of all the knee joints, making the machine crawl sideways like a crab.

The human driver of the machine has three kinds of control. First, he can regulate the amount of oil flowing in the system because he can control the displacement of the hydraulic pumps. Separate pumps are provided for the legs on the left and right sides so that the driver can steer by making the machine crawl faster on one side than on the other. The settings of the steering controls are reported to the microprocessor so that it can position the feet properly. For example, if the machine is turning to the right, the front feet must be transferred to the right and the rear ones to the left to accommodate to the turn. If the machine is walking backward, which is achieved by reversing the flow of oil, the feet must be transferred backward with each step rather than forward. As each foot is lifted from the ground the control computer picks a target position for it based on the current rate and direction of oil flow set by the driver. When a supporting foot nears the limit of its travel, the control computer initiates its lift and transfer to a new foothold. If any supporting foot actually reaches the limit of its travel, the microprocessor stops the vehicle until that foot can be lifted from the ground and transferred to a new foothold where it has room to move.

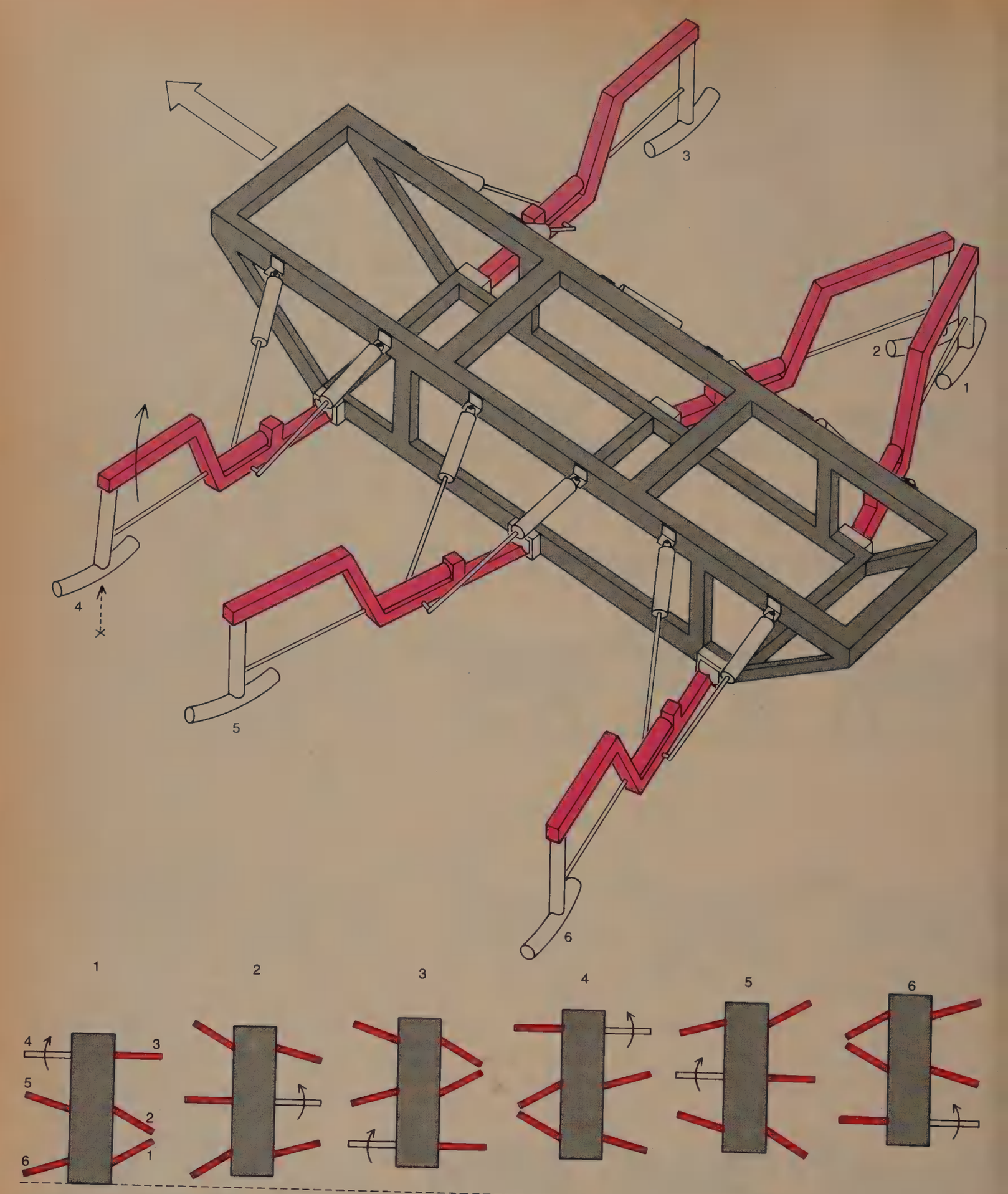
The second kind of human control of the vehicle establishes the attitude and



■ leg touching the ground, ■ open circle ■ leg in the air. The alternating-tripod gait always provides stability. In creeping with ■ wave gait the insect's adjacent legs move successively.



tographs made by Eadweard Muybridge 100 years ago that settled a long-standing debate on whether ■ horse in ■ trot leaves the ground. Many other animals also do so in running.



SIX-LEGGED MACHINE built by one of the authors (Sutherland) moves by crawling. It does not have to balance. The six legs are controlled by a built-in microcomputer. Power comes from an 18-horsepower gasoline engine that drives separate hydraulic pumps for the left and right legs. A human driver steers the machine by making the oil flow at different rates on the two sides. Sensors report the driver's commands as well as the position of each leg and the forces on it to the microcomputer, which employs the information to choose the order and path of leg motion. Six legs ensure stability because at least three are always on the ground. Passive hydraulic circuits simplify the computing task; a leg that is supporting weight can either be connected to the drive unit or can coast, being pushed by the ground according to the motion generated by the other legs. The diagram at the bottom indicates the position of the legs in a walking cycle, which can be denoted as (4;2;6;3;5;1;). A solid rectangle represents a leg touching the ground and an open rectangle represents a raised leg moving forward as indicated by the arrow. In the numbered walking cycle the semicolons denote the sequential use of the machine's six legs.

ground clearance of the machine. The driver can set a control that changes the vertical support position for the left and right feet to make the vehicle roll left or right. Similarly, he can indicate different vertical support positions for front and rear feet to make the machine pitch forward or back. Another control enables him to indicate that the vertical support positions for all six legs should be raised or lowered collectively to change the ground clearance of the vehicle.

The third kind of human control will achieve careful placement of the feet for operation on very rough terrain. We have not yet decided how to provide this kind of control. A walking truck built some years ago by Ralph Mosher at the General Electric Company depended exclusively on manual control of foot placement and was therefore quite tiring to drive. We believe selection of the gait may also be important, but we have not yet had enough experience to know whether it could be done automatically or whether human inputs will be needed. It is precisely to answer such questions that we have built the machine.

The other subject of our attention is walking and running where balance plays a role. Until a century ago people still debated whether or not a horse in a trot had all its legs off the ground simultaneously. The stop-motion photography of Eadweard Muybridge settled the debate, showing that a horse does leave the ground entirely during a trot. A running person does so too, as do the dog, the cheetah and of course the kangaroo. Such animals not only walk, which requires dynamic balance, but also run, employing ballistic motions effectively to increase their rate of travel.

There are two fundamental differences between a crawling vehicle that is statically balanced and one that is dynamically balanced. The first difference is in the definition of stability. A crawling vehicle is stable if its legs provide at least a tripod of support at all times to ensure that it does not tip over; a dynamically balanced walking or running vehicle can be allowed to tip for brief intervals. Motions of the legs and the body ensure that a single tipping interval is brief and that an adequate base of support is maintained on the average. For example, a running man touches the ground alternately with his two legs, providing a base of support for his body only over time.

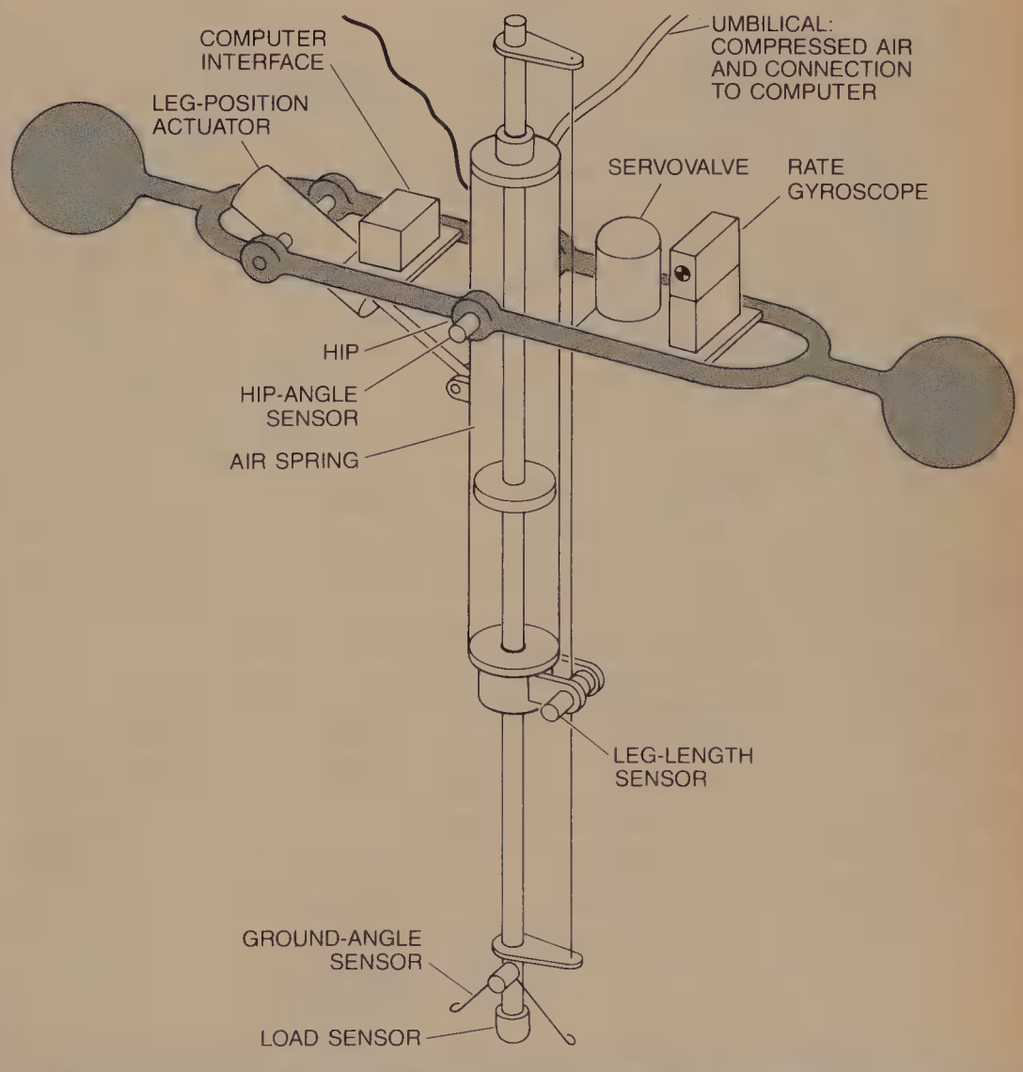
The second difference between static and dynamic balance is in the consideration of speed and momentum. Static balance assumes that the configuration of the supporting legs and the position of the center of gravity are adequate to specify stability; it ignores the vehicle's motion. Such static computations are not always sufficient. For example, a fast-moving vehicle might tip forward if it stopped suddenly with the center of gravity too close to the front legs. In order to understand the greater mobility of walking and running systems one must both relax the definition of stability and account for velocity in computing balance.

It is to study the problem of balance in its simplest form that one of us (Raibert) and his co-workers at Carnegie-Mellon University have built and demonstrated a machine that hops on its single leg and runs like a kangaroo, in a series of leaps. The device can be thought of as a computer-controlled pogo stick. We have been encouraged by the remarkable simplicity of the balancing algorithm. In its present form the machine is limited to movement in a single plane, so that it can tip over in only one direction.

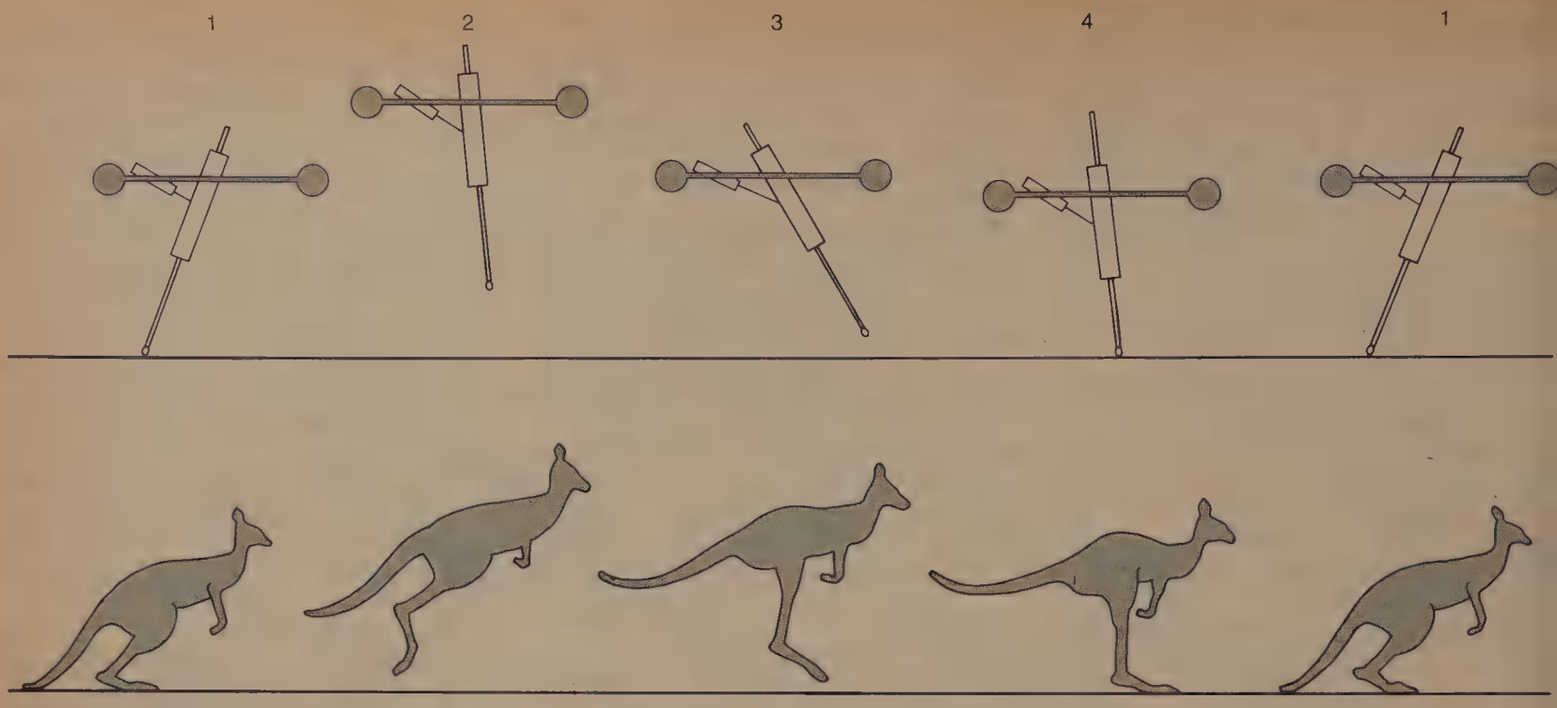
The machine has two main parts: a body and a leg. The body provides the main structure and carries valves, sensors and electronics. The leg is a simple mechanism that not only changes length along its axis but also pivots with respect to the body at a hinge called the hip. The leg bounces on a spring with adjustable tension, much like a human leg with its springy muscles and tendons. The spring is an air cylinder in which pressures are controlled with sensors and valves. At the bottom of the leg is a small foot.

The pivoting motion of the leg is controlled by a second air-operated actuator that applies torques at the hip hinge. A simple on-off valve controls the leg spring, but control of the pivot angle of the leg requires a proportional servovalve, that is, a feedback device that responds in proportion to the strength of the signal it receives. Because the moment of inertia of the leg is less than 10 percent of the body's moment of inertia the leg can pivot during flight without imparting much motion to the body. The tilt of the body is measured by a gyroscope, enabling the control computer to maintain the body in a level attitude. Other sensors measure the angle of the hip, the length of the leg, the air pressure in the leg spring, the angle between the leg and the ground and the force of the leg's contact with the ground.

Three separate servo-control loops



TWO-DIMENSIONAL HOPPING MACHINE serves to study the problems of controlling motion with balance. The machine normally operates while leaning parallel to a tilted wall, separated from it by a cushion of air; the machine can tip only in the plane defined by the wall. A computer receives data from the angle sensors, a pressure sensor and a switch on the foot. It controls hopping by adjusting the pressure in the pneumatic chamber that functions as a spring. The computer also applies torque between the leg and the body to regulate the angle of the hip. This angle determines the horizontal displacement of the foot and therefore influences balance.



HOPPER IN MOTION operates cyclically, as all legged systems do; the leg alternates between periods of support and periods of flight. At the left the machine is about to begin a leap. While it is in the air the leg swings forward at the hip in preparation for the next landing. At touchdown the leg spring shortens to its minimum length to provide for the next leap. A ground-contact sensor acts as a trigger for the vertical-control program. The machine also has feedback loops to control attitude and balance in synchrony with the vertical control. Like a pogo stick, the machine can balance only while it hops. The hop is like the kangaroo's movement Muybridge called a ricochet.

regulate the machine. One loop controls vertical motion, one balance and one body attitude. Each loop is synchronized with the basic hopping motion.

The first loop controls the height of the hopping motion. It adds or removes energy from the motion in order to achieve the correct hopping height and makes up for the energy lost during each hop. The height control does both tasks by periodically adding air to or releasing it from the main drive cylinder to adjust the effective tension of the air spring. In other words, the height control governs the timing and the magnitude of the power delivered to the hopping drive mechanism, thereby achieving the desired hopping height. When a desired hopping height has been achieved, most of the energy needed for the next hop is recovered from the spring, in which it was stored during the previous landing. As long as the hopping motions are relatively stable the task of managing the hopping energy of the machine is not particularly difficult.

The second servo-control loop provides for the balance of the machine by positioning the foot while the machine is in flight so that the next landing is made in a balanced posture. The calculation of the correct foot position takes into account both the forward speed of the vehicle and the inclination of the body. A single computer algorithm for balance works when the machine is hopping in place, accelerating to a run, running at a constant velocity, leaping over objects and slowing to a stationary hop.

When the machine is hopping in place, the leg and foot are moved small distances to compensate for external disturbances and the errors of previous hops. When the machine is to start running, say to the right, the foot is moved first to the left to unbalance the vehicle so that it starts to tip in the desired direction. Stable running is just like hopping in place except that the balancing adjustments supplement large sweeping motions of the leg, which are determined by the rate of travel. Stopping is much like starting except that the machine is made to tip in the direction opposite to the direction of movement.

The third servo-control loop stabilizes the attitude of the body to keep it upright. It provides torques between the leg and the body while the foot is on the ground in order to achieve the desired attitude during the next flight. The effectiveness of this servo depends on good traction between the foot and the ground. The attitude servo that operates when the foot is on the ground shares the hip-drive mechanism employed by the balance servo that operates while the machine is in flight. Certain subtle details of the change from one control mode to the other are associated with detecting the start and finish of each flight. The torquing mechanism must be idle during these events lest the foot slip on the ground.

When an animal runs, its legs swing back and forth through large angles to provide balance and forward drive. We have found that such swinging motions of the leg do not have to be explicitly programmed for a machine but are a natural outcome of the interactions between the controllers for balance and attitude. Suppose the vehicle is traveling at a constant horizontal rate and is landing with its body upright. What must the attitude controller do during stance to maintain the upright attitude? It must make sure that no torques are generated at the hip. Since the foot is fixed on the ground during stance, the leg must sweep back through an angle in order to guarantee that the torque on the hip will be zero while the body moves forward.

On the other hand, what must the balance servo do during flight to maintain balance? Since the foot must spend about as much time in front of the vehicle's center of gravity as behind it, the rate of travel and the duration of stance dictate a forward foot position for landing that will place the foot in a suitable spot for the next stance period. Thus during each flight the leg must swing forward under the direction of the balance servo, and during each stance it must sweep backward under the control of the attitude servo; the forward and back sweeping motions required for running are obtained automatically from the interplay of the servo-control loops for balance and attitude.

We are now building a version of the machine that will balance in three dimensions and therefore be able to move around on an open floor. We have written and tested a computer simulation of the motion of such a machine and have found that control in three dimensions can be broken down into the same three servo-control loops we have described.

Our work in making the one-legged machine run was greatly aided by a

thought that came to us as we went along. It was that running can best be understood by breaking it down into the three parts we have discussed here: height control, balance control and attitude control. Partitioning the control into these three parts has made the complex behavior of legs in walking and running much easier to understand. This insight has led to a fairly simple control system that makes the one-legged machine balance and run.

Our success in this effort encourages us to think about building dynamic-motion machines with more than one leg. We believe the right way to think about such machines is to focus first on their up-and-down and balancing behavior, postponing the complications introduced by forward motion. The notion that hopping is the main activity was natural to the one-legged machine and provided an effective way to think about its behavior, but it seems less natural for machines with several legs. Perhaps it seems less natural because we are accustomed to seeing animals run and want to understand their behavior all at once.

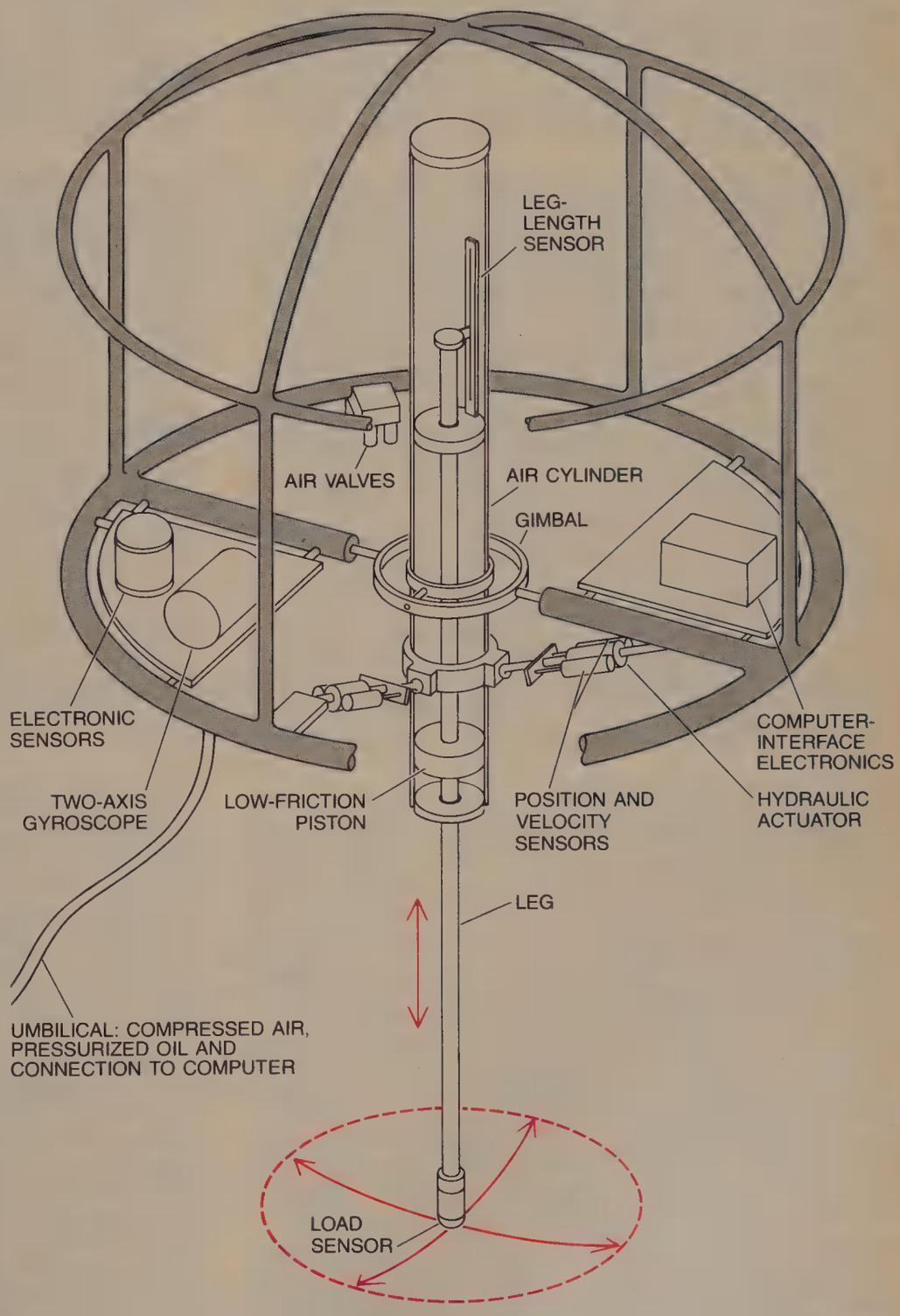
A four-legged machine hopping in place might use any of several sequences of leg activity. The simplest pattern would be to hop simultaneously on all four legs. It is not hard to imagine that the same three servomechanisms that control the one-legged hopping machine might control the motion of a four-legged vehicle in this mode. In fact, the attitude-control loop that keeps the body upright might be substantially simpler because of the broader base of support. When the machine moved forward, the legs would swing together in a pattern of motion that could easily be generated by the same control mechanisms as those that serve the one-legged machine.

Another possible gait for a four-legged machine hopping in one place is bouncing on diagonally opposite pairs of feet. Again it does not stretch the imagination too much to see how one might separate the control of such a vehicle into a height control, a balance control and a body-attitude control. The height control would add energy to the hopping motion to keep the hopping height at the desired level. The balance control would position the two raised legs in such a way as to maintain balance. The body-attitude control would apply appropriate torque to the pair of legs on the ground. The attitude and balance controls would alternate in the use of the same leg actuators, as they do in the one-legged machine. Moreover, just as we have found in that machine, forward motion could easily be accommodated by moving each raised foot to a forward position chosen to make the average balance force of the leg zero during the next stance period. The resulting motion is a trot of the kind common among four-legged animals.

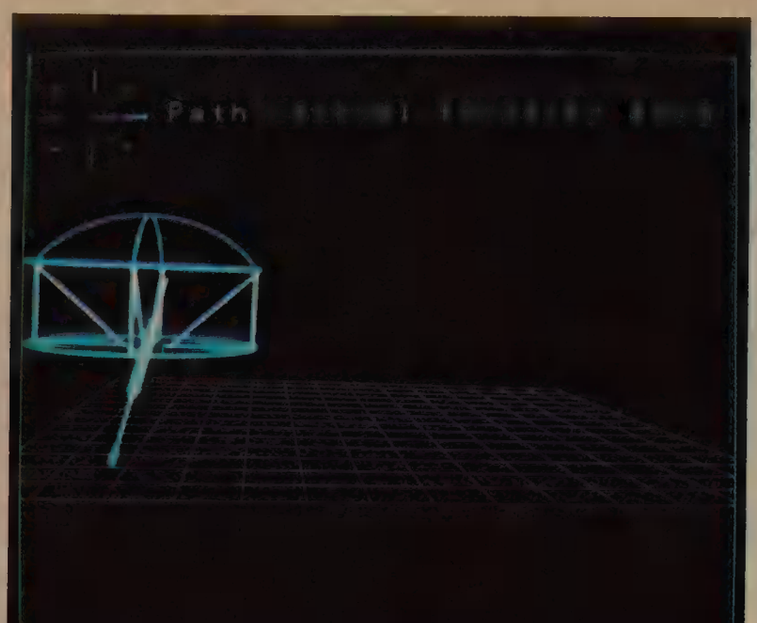
Two other gaits can similarly be understood by separating the control of each leg into vertical, balance and body-attitude components. In the gallop the rear legs land slightly sooner than the front ones. The body attitude is allowed to change during flight so that a nose-up attitude is seen as the rear legs touch down and a nose-down attitude develops as the front legs take off. The bound is a variation of the gallop in which the front legs operate nearly in unison and so do the rear legs but the front and rear actions are equally spaced in time. It is the bound that enables the cheetah to sprint at speeds of more than 60 miles per hour.

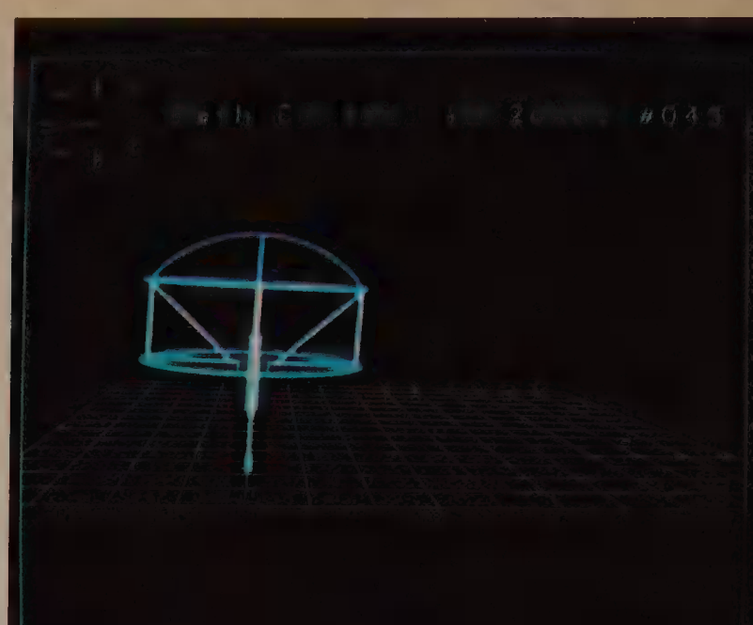
Efficient motion over the ground requires that little energy be lost during each motion of the machine. We have already mentioned how the vertical motions of legs can be made efficient by storing energy in elastic elements. What about fore-and-aft leg motions?

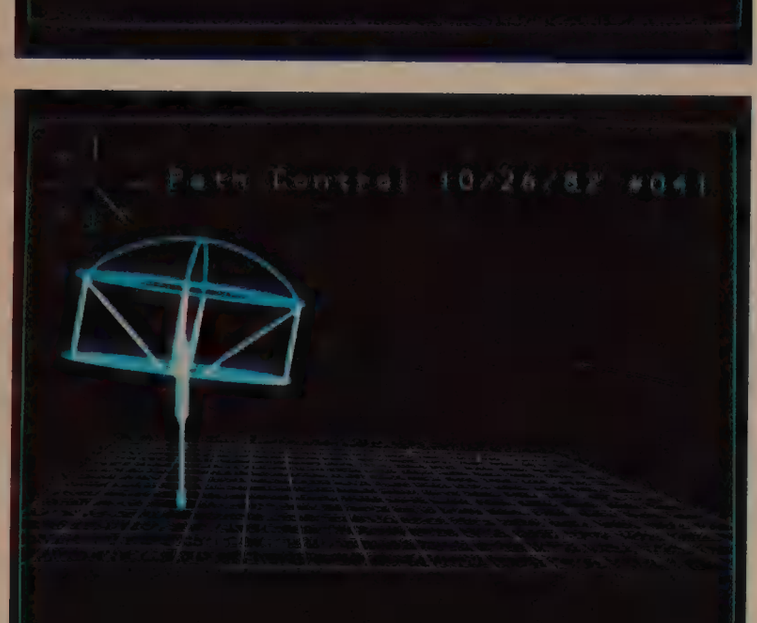
At high speeds over the ground the legs of a vehicle will have to move forward and back quite rapidly. Most of the energy expended by a running animal goes into generating these leg mo-

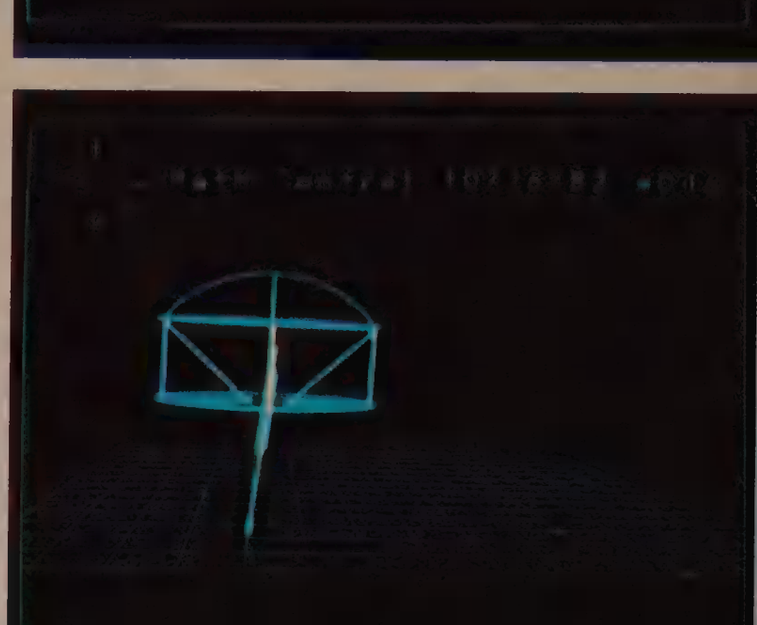


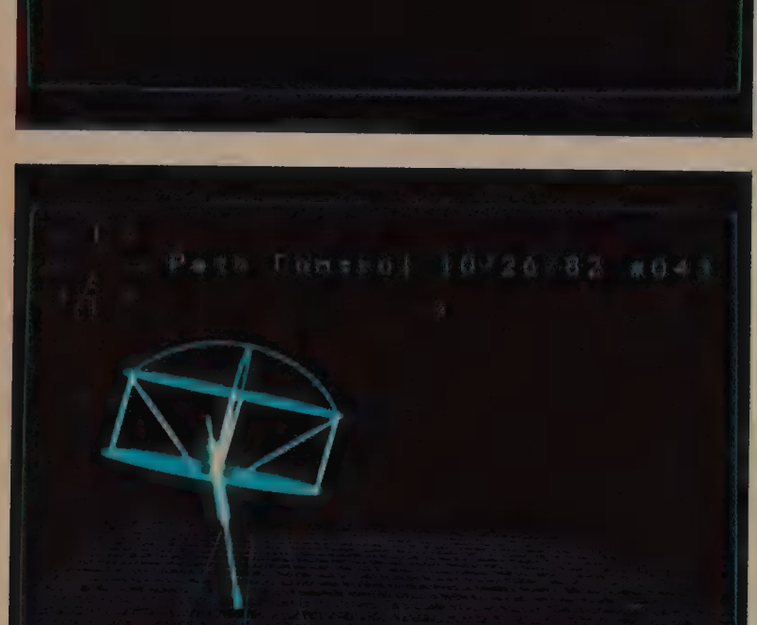
ADVANCED HOPPING MACHINE now under development is designed to operate in three dimensions. It is about one meter high, weighs 20 kilograms and is connected to a nearby computer. Compressed air provides the hopping power and regulates the height of hopping. The actuators that position the foot are hydraulic. During flight they position the foot to maintain balance. When the foot is on the ground, they maintain the machine in an upright body posture.

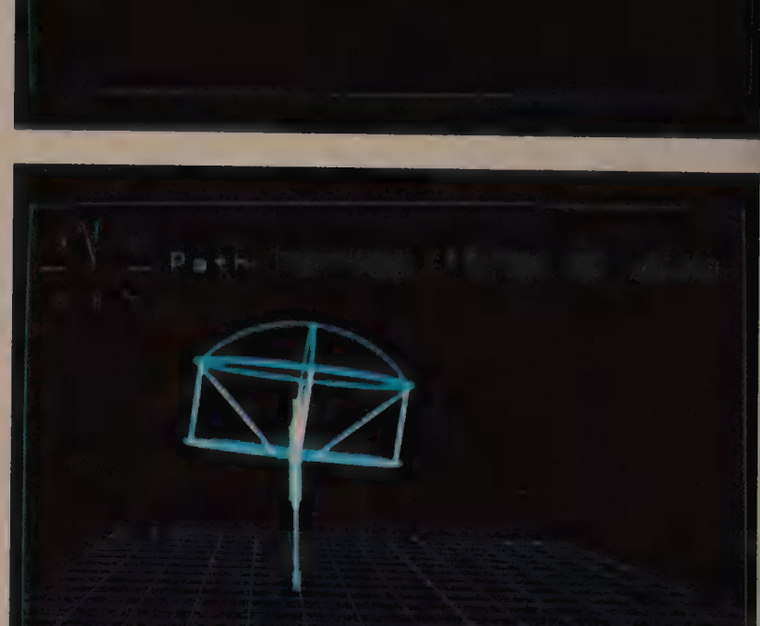












SIMULATED MOTION of the three-dimensional hopper is shown in these photographs from the display of the computer that worked out the motion. Here over a period of approximately .7 second the machine is shown balancing itself while it lands and takes off again.

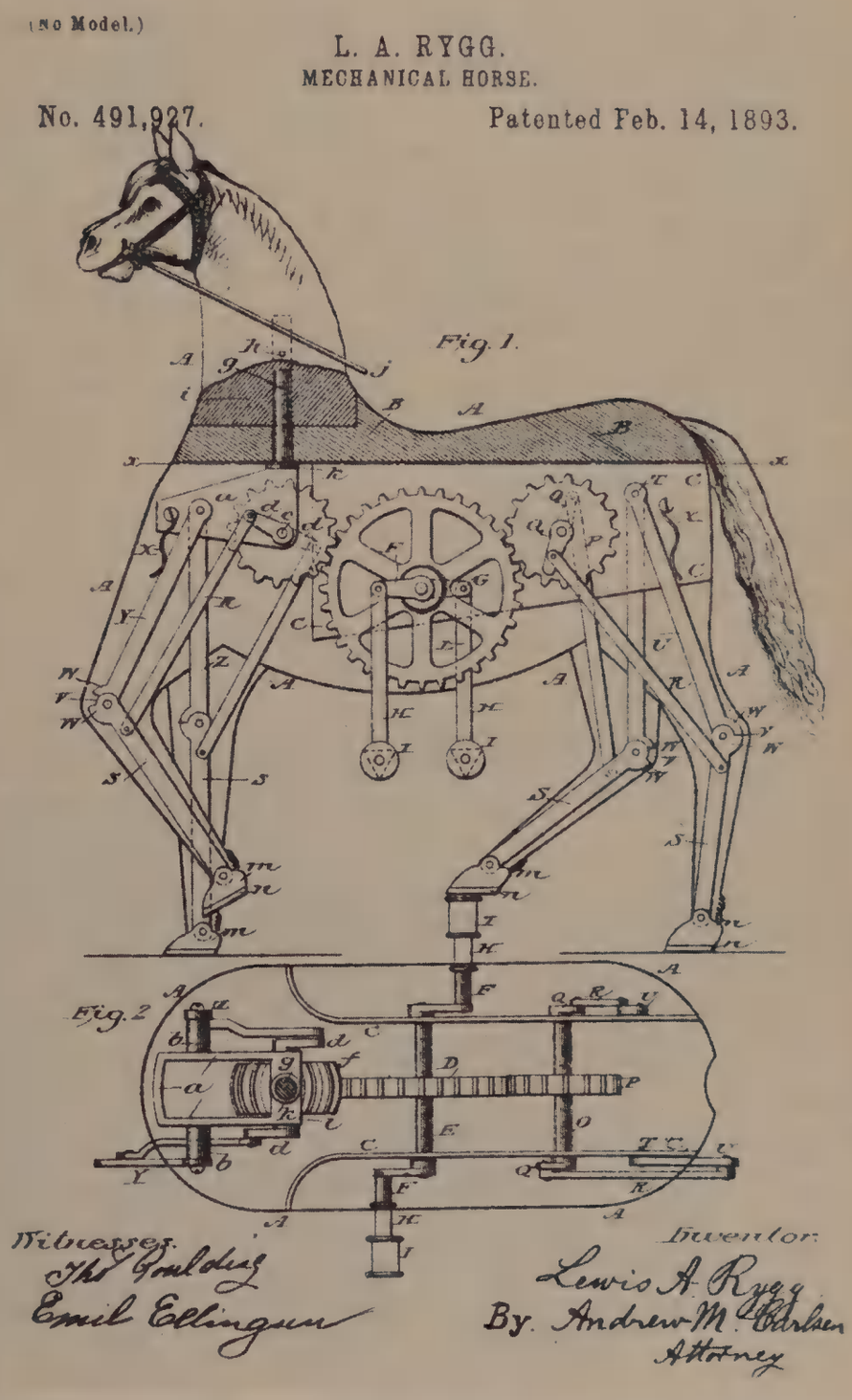
tions. In our one-legged machine these motions are provided by a conventional proportional servomechanism. In such a system the kinetic energy of the leg as it swings is entirely lost as the leg is brought to rest momentarily at each extreme angle. The hopping motion, on the other hand, is obtained by a self-resonant system made up of the leg spring and the mass of the machine, so that the height servo need only add or subtract a small amount of energy to maintain the hopping height. It is obvious that if a machine with several legs is to be made efficient as well as fast, it will have to incorporate some kind of self-resonant system for the fore-and-aft motion of the legs as well. One might design such a mechanism with springs between the legs to make the legs oscillate like a tuning fork at a frequency appropriate to the vertical bouncing rate.

Although we believe we understand how to build a four-legged machine that can run with any of the common gaits we have described, there remain many interesting questions associated with starting and stopping such a machine and selecting its gait. We can easily see how to start forward if the machine is already hopping in place. What is much less obvious is how to coordinate the transition from a standing start to full-speed running. Similarly, how and when should such a machine change from one gait to another? A running horse switches its lead as it turns, that is, it changes which of the two front legs slightly precedes the other. What computations should be done to make such minor changes in the pattern of leg motion? We find these problems fascinating, both as engineering questions in the form "What should we build?" and as scientific questions in the form "How do living systems work?"

A much more difficult problem is how to choose footholds for the machine. The function of vision in walking and running by people and animals, particularly the ability to choose sensible places to put the feet, is not well understood. One can imagine avoiding this problem by having the machine run fast only over smooth ground and by having some kind of human assistance to choose a safe path. It will probably be desirable to scan the ground ahead of the vehicle for holes. Still, just as a galloping horse runs the risk of stepping into a gopher hole, so we must expect that a running machine will also get into that kind of trouble.

The mobility of off-road vehicles is limited by two factors. First, the continuous footprint of wheels and tracks prevents wheeled and tracked vehicles from making use of the discontinuous points of support that are available to a legged vehicle. We are encouraged to think that the state of the computer art is now sufficiently advanced to allow the construction of adequate control systems for legged vehicles, and so the legged alternative for high-mobility vehicles can be seriously considered. Indeed, the Defense Advanced Research Projects Agency is already sponsoring research on such vehicles, including partial support for our projects.

A second source of mobility is narrow width: a motorcycle can get into places a jeep cannot reach. Narrow legged vehicles can be built, but they will have to balance, at least in the sideways direction. Involved as we are both in the construction of a six-legged vehicle that can crawl without attention to balance and in studies of walking and running with balance, we believe the effort to understand balance is much more important. We think the experiments with six-legged crawlers now under way in our laboratory and elsewhere are mainly exercises in the control of multiple legs and are not in themselves useful; such crawlers will ultimately be replaced by machines with fewer legs that can balance. Mastery of balance will be the key to building high-mobility machines that walk and run.



MECHANICAL HORSE was the subject of a patent obtained by Lewis A. Rygg in 1893. The lower drawing is a plan view along the line *x–x* of *Fig. 1*. The stirrups doubled as pedals that were to enable the rider to power stepping motions. Steering was to have been done with reins that moved the head and forelegs from side to side. Apparently the machine was never built. It would have been similar to many modern walking toys. Since they have no sensing or computing facilities, they cannot adapt to variations in terrain. They only crawl on flat surfaces.

The Hair Cells of the Inner Ear

They are exquisitely sensitive transducers that in human beings mediate the senses of hearing and balance. A tiny force applied to the top of the cell produces an electrical signal at the bottom

by A. J. Hudspeth

The sense of hearing, the sense of balance that enables human beings to walk upright, the ability of certain animals to detect vibrations of the ground and the ability of fishes to detect the displacement of water would appear to have little to do with one another. Actually the four senses are closely related. Indeed, each of them is made possible by the same sensory receptor. The receptor is called a hair cell and is named for the hair bundle, a group of fine projections that extend from its upper surface. The hair cell is an extremely sensitive mechano-electric transducer, that is, it converts a mechanical force into an electrical signal. The mechanical force is the stimulus applied to the hair bundle; the electrical signal is the message relayed to the brain.

Each hair cell is sensitive to only a limited range of stimuli. Therefore if the organism is to obtain useful information about its environment and about its own movements the output from thousands of receptors must be combined. The requisite number of hair cells are found in several small sensory organs in the inner ear. In human beings arrays made up of thousands of hair cells with slightly different sensitivities are found in six organs in each ear. The combined responses of the cells yield information about linear acceleration in any direction, about angular acceleration along three perpendicular axes and about acoustic tones with a wide range of frequencies.

The general structure of the hair cell and the receptor's sensory role have been known for many years. How the individual hair cell works, however, is only now becoming clear. To examine events on the cellular scale I have worked with single hair cells from the inner ear of the bullfrog. A microscopic probe is employed to push the bundle back and forth as an electrode records the cell's electrical output. With this setup I have obtained the first recordings made from single hair cells during the application of a precisely controlled mechanical stimulus directly to the hair bundle.

Each hair cell can respond to remarkably small stimuli. Hair cells from mammals begin to respond when the tip of the hair bundle is moved no more than 100 picometers (trillionths of a meter); this distance is about the same as the diameter of some atoms. In addition, recordings from isolated hair cells show that each receptor has one direction of maximum sensitivity. When the hair bundle is displaced in any one direction, the cell responds only to the component of motion that is in the direction of maximum sensitivity. Along with light microscopy and electron microscopy recordings from individual hair cells are beginning to yield a rich understanding of the workings of the sensory receptors of the inner ear. Many questions nonetheless remain, among the most intriguing of which is: What is the exact molecular mechanism whereby the displacement of the hair bundle changes the electrical properties of the cell and causes a message to be sent to the brain?

The hair cell is generally either cylindrical or flask-shaped and is always found as part of an epithelium, a sheet of cells that is at most a few cells thick. It is closely related to the neuron, or nerve cell, although it does not have axons or dendrites, the fibers that extend from neurons and transmit electrical signals in the nervous system. For this reason the hair cell is sometimes referred to as a paraneuron. In the epithelium where the hair cell is found its apical, or upper, end lies flush with the upper ends of the supporting cells that surround it. The apical ends of the hair cells and those of the supporting cells form a smooth surface above which the hair bundle extends. In different cells the length of the hair bundle ranges from about three micrometers (millionths of a meter) to more than 100 micrometers.

The detailed structure of the hair bundle differs among the hair cells of different species and even among hair cells from different organs of the same individual. The basic form of the hair bundle, however, is the same in all vertebrate animals. Each bundle consists of from 30 to 150 thin, rod-shaped extensions. The extensions are of two kinds; they are both called cilia but nonetheless have quite different internal structures. All but one of the extensions are called stereocilia. Stereocilia are cylindrical or club-shaped organelles with a core made up of tightly packed filaments. The plasmalemma, the surface membrane of the cell, extends over the filamentous core of the stereocilium much as the finger of a glove covers the human finger. The diameter of the stereocilium ranges in different cells from about .2 micrometer to one micrometer.

In spite of the term, stereocilia are not true cilia. A true cilium, such as the cilium of a sperm cell, has a complex and highly differentiated central structure called the axoneme. Most true cilia are capable of an independent motion much like that of an oar in rowing; the motion originates in the axoneme. The hair bundle has one true cilium: the kinocilium. It is about .25 micrometer in diameter. Like the axoneme of other true cilia, the axoneme of the kinocilium consists of two central tubules surrounded by nine "doublet" microtubules, pairs of small tubes that share a central wall. The kinocilium is not, however, capable of independent motion in the hair bundle. At the end away from the main body of the cell it is usually attached to the adjacent stereocilia.

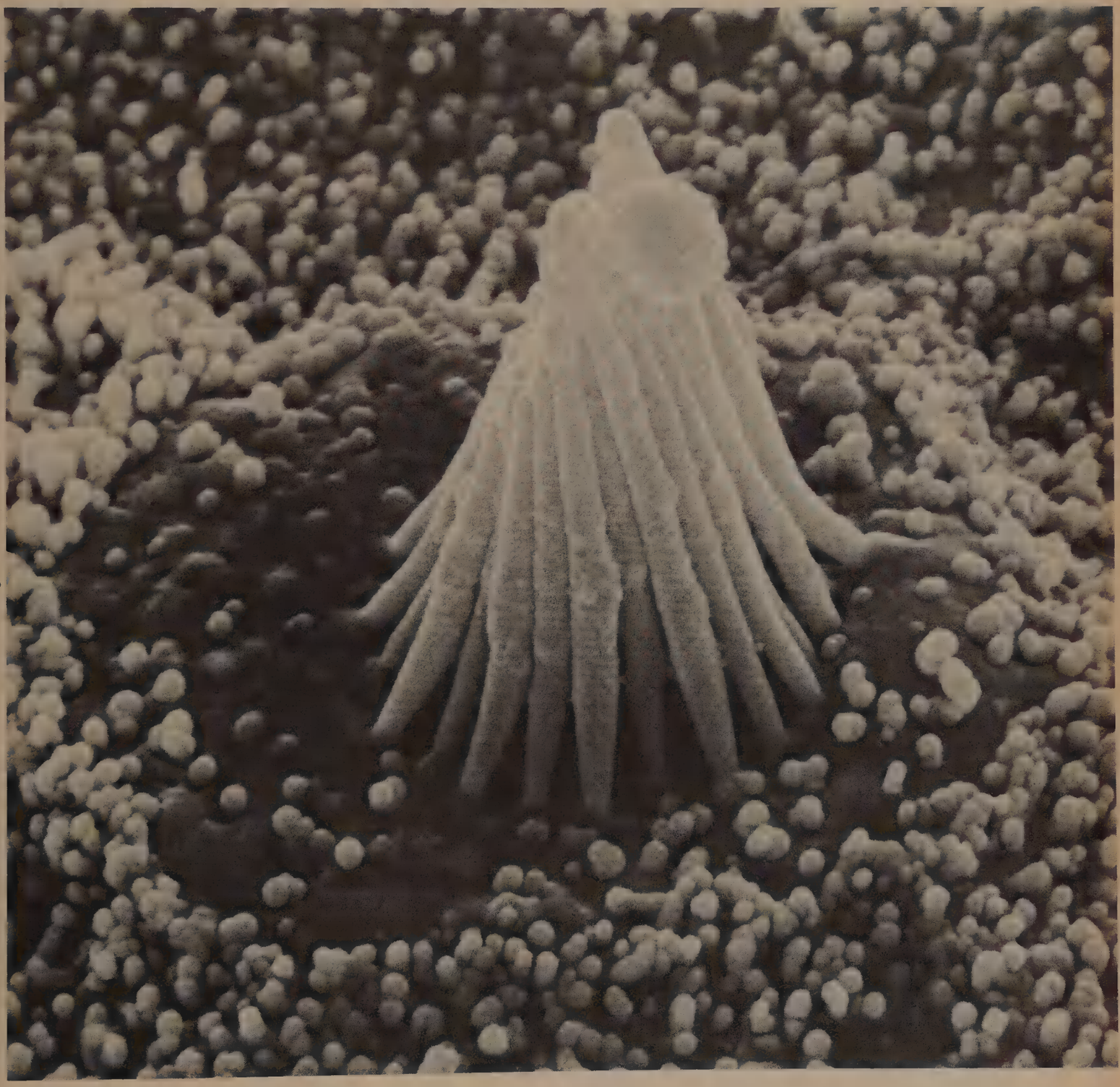
The arrangement of the cilia in the hair bundle is highly regular. The stereocilia are in a hexagonal array: each one is surrounded by six others that are equidistant from it. The stereocilia are not of equal length. The hair bundle is circular in cross section and along one diameter of the circle there is a progressive increase in the length of the stereocilia. Along any axis perpendicular to that diameter, however, the cilia are the same length. Thus the hair bundle has ranks of cilia of equal length that are arranged in order of size. As a result of this arrangement the hair bundle has a plane of bilateral symmetry. In the intact cell the stereocilia are not uniform-

ly perpendicular to the apical surface. Instead they lean together to form an obliquely truncated cone, with the kinocilium lying on the edge of the cone in the plane of bilateral symmetry. Near their tops the cilia approach one another closely or indeed touch.

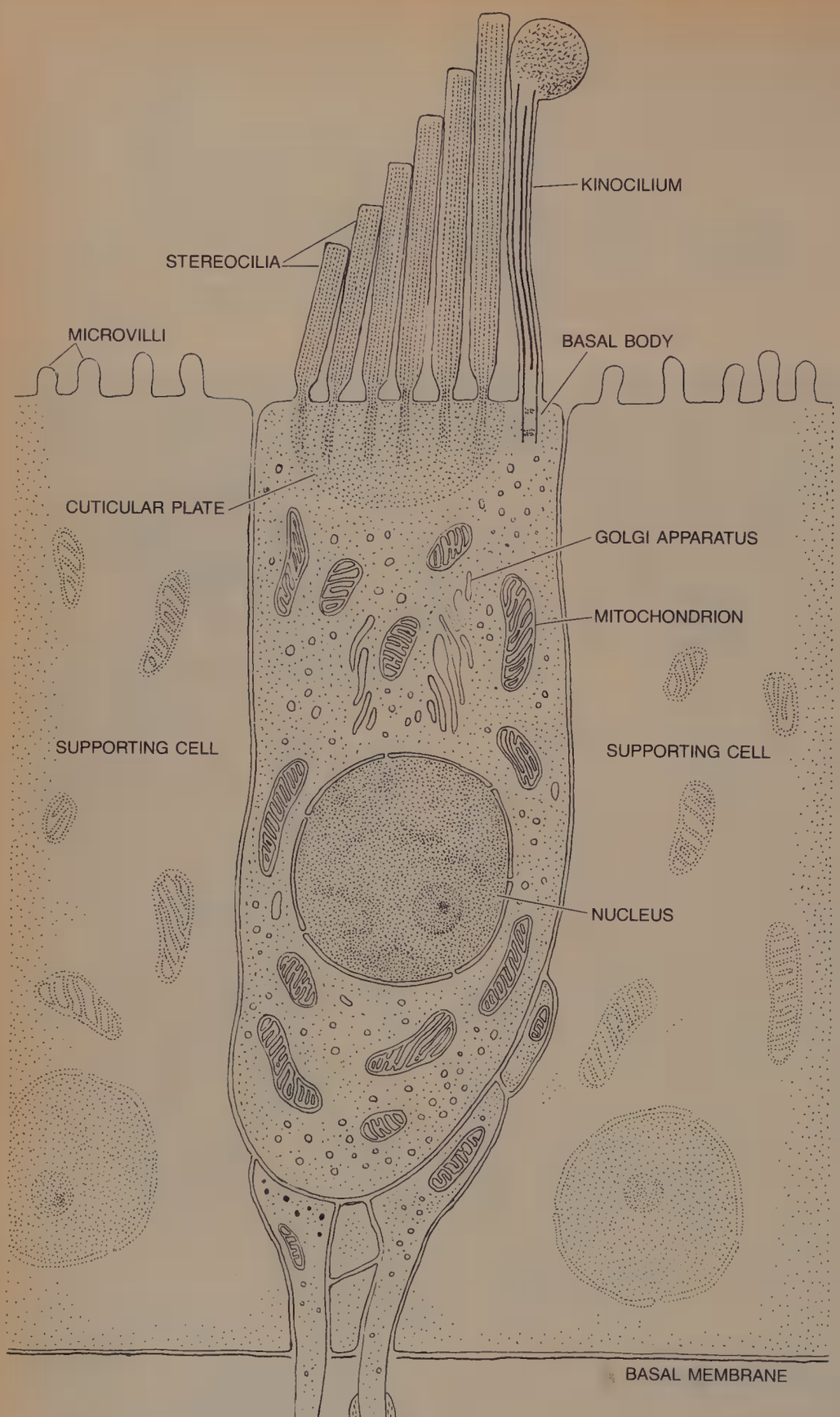
In the growth of a human embryo the cells from which the inner ears develop first become recognizable as two placodes, or thickened patches of cells. The placodes are on the surface of the embryo on each side of the developing brain stem. As they increase in size they buckle inward and ultimately lose their connection with the surface of the embryo. Once they are separated from the surface they are known as otic vesicles. Each otic vesicle is a minute pouch of epithelial tissue filled with fluid.

The otic vesicle undergoes a complex development that includes growth, bending and the fusion of its parts. The result of the process is the formation of the six receptor organs in which hair cells are found. The three semicircular canals, the receptor organs for angular acceleration, are toroidal (doughnut-shaped) tubes. They are joined together in such a way that when the head is in its usual orientation one canal lies in the *x* plane of three-dimensional space, one in the *y* plane and one in the *z* plane. The utricle and the saccule, the organs that are sensitive to linear acceleration, are small, irregular pouches. The cochlea, the organ of hearing, has a spiral form.

The fully formed inner ear has a com-



HAIR BUNDLE is the structure for which the hair cell is named; it is responsible for the cell's mechanical sensitivity. The bundle is enlarged 14,000 diameters in this scanning electron micrograph of a hair cell from the inner ear of a bullfrog, made by Richard A. Jacobs of the California Institute of Technology. Hair cells are found in epithelia, or sheets of tissue a few cells thick, where they are surrounded by supporting cells. The hair bundle shown here consists of some 50 fine fibrous extensions. All but one of the extensions are stereocilia, which in spite of their name are not true cilia. The hair bundle also has one true cilium, the kinocilium, which ends in the bulbous swelling at the upper-right edge of the bundle. The structure is shown in the resting position, where it forms an upright, obliquely truncated cone. Bending the hair bundle away from the resting position causes the hair cell to relay an electrical signal to the brain. The many small light-colored projections around the hair bundle are microvilli extending from the apical surfaces of the nearby supporting cells.



HAIR CELL is shown in cross section. The cell is generally either cylindrical or flask-shaped. Although they are closely related to nerve cells, hair cells have no dendrites or axons. The organization of the hair bundle varies among species in its details but its overall structure is the same in all vertebrate organisms. Along one axis there is a progressive increase in the length of the stereocilia. Along the perpendicular axes, however, the stereocilia have the same length. As ■ result the hair bundle has ■ plane of bilateral symmetry; the illustration is a section through this plane. The kinocilium lies in the plane of symmetry at the tall edge of the cone. The afferent nerve ending at the basal surface of the cell transmits the receptor's response to the brain.

plex geometry that arises from the bending, twisting and fusion of one otic vesicle. For this reason the structure is referred to as the labyrinth. In spite of the complex form of the labyrinth, each of the six sensory organs has a fundamental simplicity of design that is derived from the form of the otic vesicle. Each organ consists of a continuous epithelial sheet supported by connective tissue. The epithelium forms a closed space filled with the fluid called endolymph. The space outside the epithelium is filled with ordinary extracellular fluid. Where substantial volumes of the extracellular fluid accumulate in the inner ear the fluid is termed perilymph. The entire structure is encased in cartilage and bone.

Thus the epithelium separates two different fluids. Embedded in the epithelium, the hair cell is in contact with both of them. The apical surface of the cell is in contact with the endolymph in the interior of the organ. The basal, or bottom, surface of the cell is in contact with the perilymph that surrounds the organ. As we shall see, these fluid compartments have an important role in the function of the hair cell.

In order to appreciate the action of the hair cell it is necessary to consider the mechanical operation of each of the organs of the labyrinth. The utricle and the saccule are positioned so that the ensemble of four organs in the two ears provides sensitivity to linear acceleration along any axis. Even more important is the fact that the paired utricles and saccules constantly monitor the acceleration due to gravity; they are crucially important in the control of posture, gait and equilibrium. Although the directional sensitivity of the utricle and saccule are quite complex, as an approximation it is reasonable to consider that the utricle responds to horizontal acceleration and the saccule responds to vertical acceleration.

The hair cells of the utricle and the saccule are found in each organ in a crescent-shaped patch of epithelium about a millimeter across. The sheet is more or less flat and includes a few thousand hair cells. In the saccule the epithelial sheet is mainly vertical and in the utricle it is mainly horizontal. A structure called the otolithic membrane, which is made up of a feltwork of protein molecules, lies parallel to the epithelium and quite close to it. The kinocilium of each hair cell is inserted into an indentation in the otolithic membrane. Embedded in the otolithic membrane and piled on top of it are hundreds of thousands of the tiny crystals called otoconia. Each otoconium is composed of calcium carbonate, usually in the form of calcite, laid down on an organic skeleton.

The density of the otoconia is several times greater than that of the en-

dolymph filling the utricle and saccule. Therefore when the organs are accelerated, the greater inertia of the otoconia causes them to lag behind the endolymph. The lagging motion of the crystals is communicated to the hair bundle by the otolithic membrane. The stimulus that actually reaches the hair cell is a force displacing the tip of the hair bundle in the direction opposite to that of the acceleration.

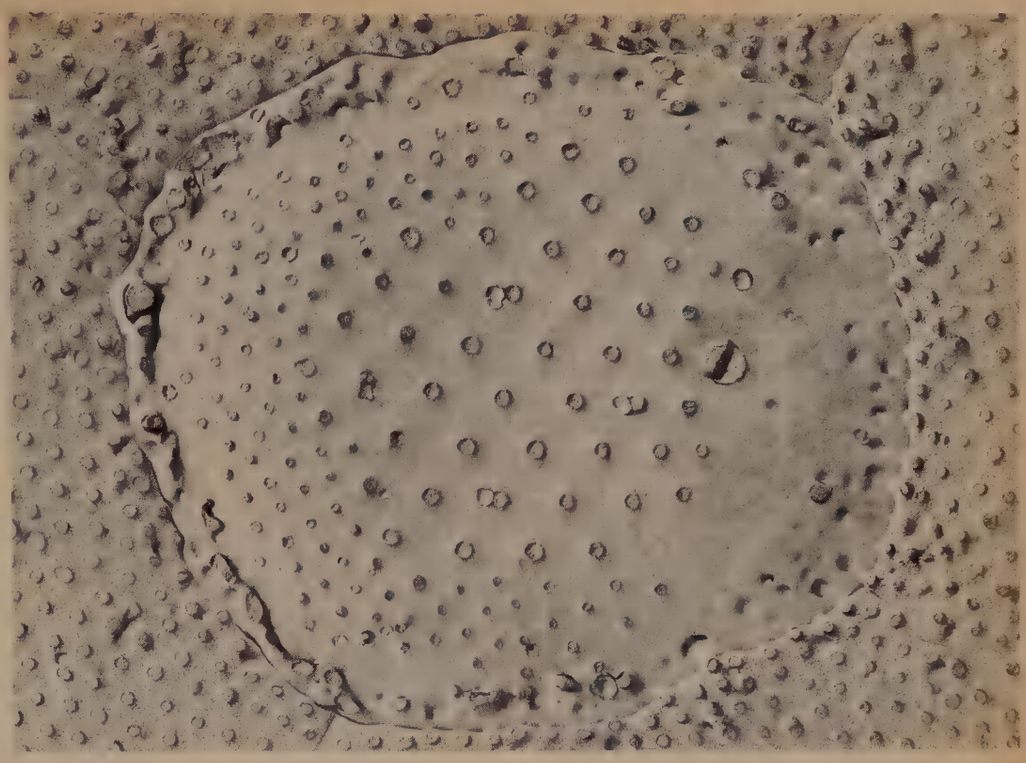
In the semicircular canals a similar mechanical force is transmitted to the tip of the hair bundle in a different way. Each semicircular canal is a tube bent in the form of a circle with a diameter of about six millimeters. In one region the cross-sectional area of the tube is larger than it is in the rest of the canal; the enlarged region is called the ampulla. In the ampulla a septum, or sheet of tissue, crosses the tube. The septum consists of a ridge of cells called the crista with a sheet of extracellular material, the cupula, extending from it. The hair cells lie in the crista. The hair bundles stretch upward and are inserted into the gelatinous cupula.

When the semicircular canal undergoes an angular acceleration (for example when the head is rotated quickly), the inertia of the endolymph in the canal causes the fluid to lag behind the walls of the canal, which move with the head. The fluid presses on the cupula, which is soft and easily distorted; the deflection of the cupula causes the hair bundles inserted into it to bend at their site of attachment to the hair cells.

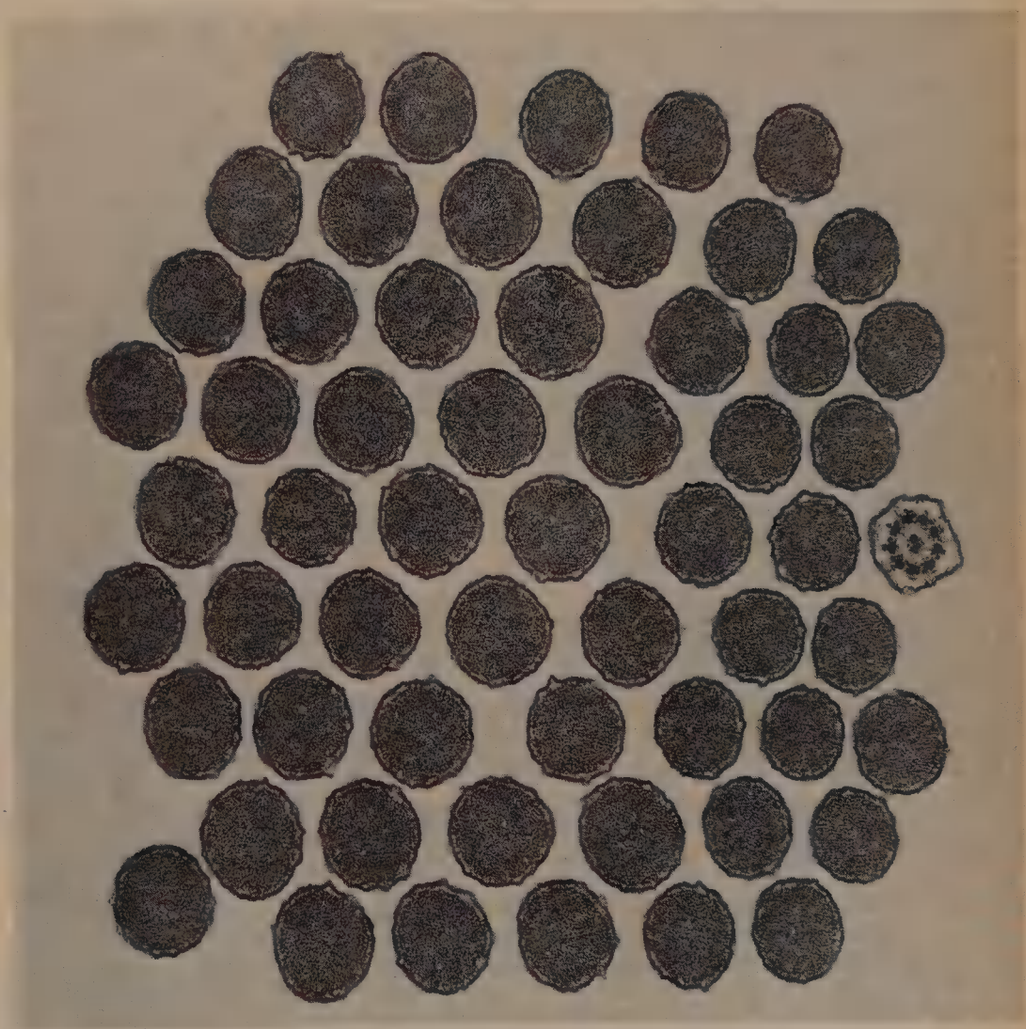
The cochlea, the sixth organ of the inner ear, has the form of a tube twisted into a spiral. The 15,000 or so hair cells in the cochlea of the human ear are sensitive to sounds with frequencies ranging from a few tens of hertz (cycles per second) to about 20 kilohertz. The hair cells are in an orderly array resting on the thin partition known as the basilar membrane. The basilar membrane winds within the cochlea and follows its spiral form. In the cochlea natural sounds with many frequencies are broken down into constituent frequencies; each constituent excites a small fraction of the 15,000 hair cells at one position along the basilar membrane.

The operation of the cochlea is too complex to be described in detail here. Moreover, its mechanism is not fully understood and is the subject of considerable controversy. As in the utricle, the saccule and the semicircular canals, however, the stimulus that reaches the hair cell is thought to be a displacement of the hair bundle. Here it will be sufficient to consider how the oscillatory pressure differences of the sound waves outside the ear are converted into a force applied to the hair bundle.

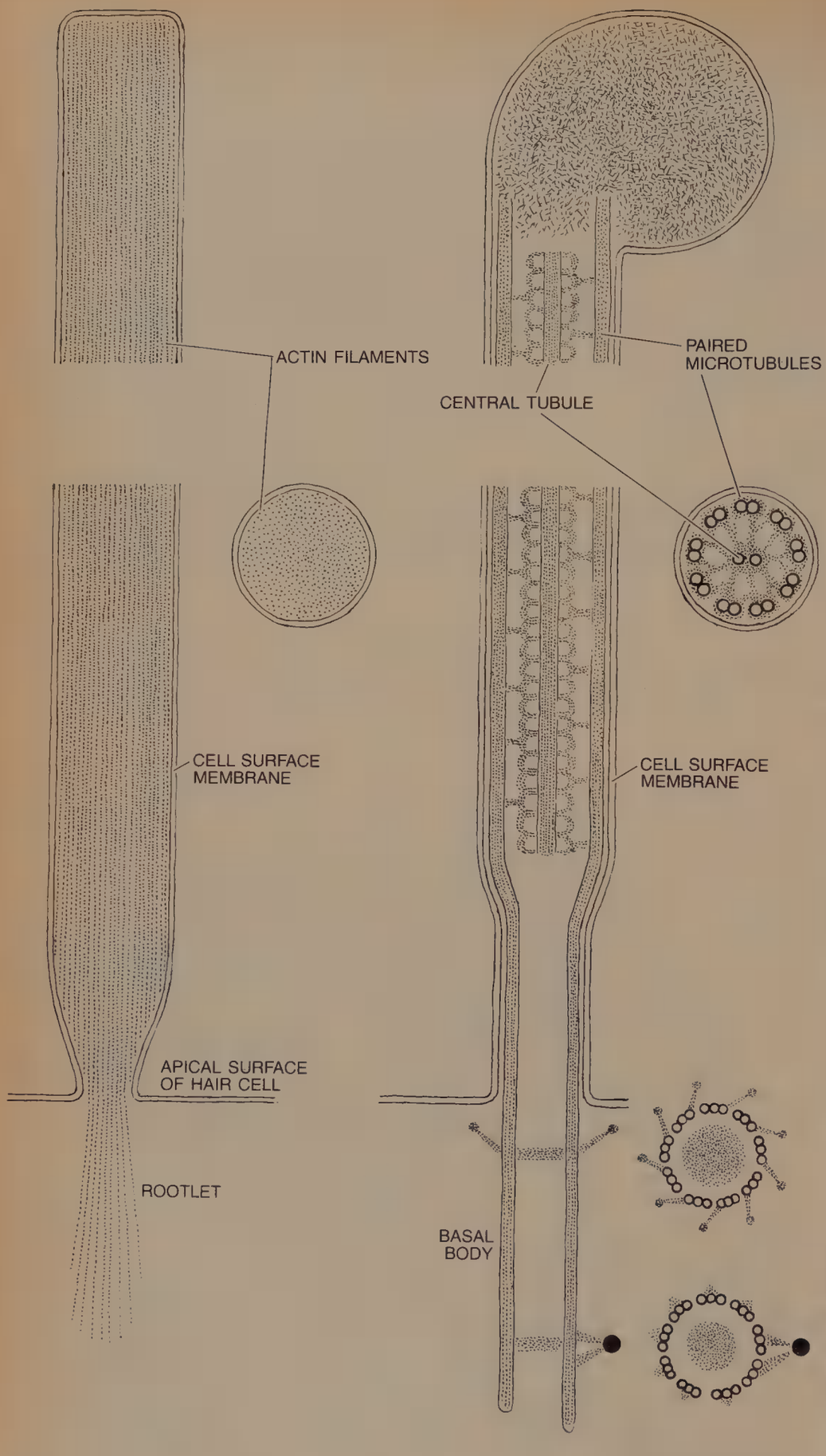
Sound waves arriving at the outer ear impinge on the membranous tympanum: the eardrum. The movement



HEXAGONAL ARRAY of stereocilia is shown in an electron micrograph of a hair cell from the inner ear of a bullfrog. The cell is enlarged 16,000 times. The micrograph was made by freezing the epithelium and shearing off everything protruding above the apical surface. The pits on the hair cell are points of insertion of the stereocilia: each one is surrounded by six others that are equidistant from it. The large broken stub at right is the stub of the kinocilium.



INTERNAL STRUCTURE of the kinocilium is quite different from that of the stereocilia. The cores of both kinds of cilia are shown enlarged 40,000 times. The core of the stereocilium is a bundle of threadlike filaments. The kinocilium is at the right. Its core is a complex structure known as the axoneme. The transmission electron micrograph was made by treating the cilia with preservatives that are better taken up by filaments and tubules than by other structures.



STEREOCILIUM AND KINOCILIUM are shown in vertical and horizontal sections. The stereocilium is at the left and the kinocilium is at the right. Each stereocilium is shaped like a pencil or a baseball bat. Its diameter ranges from .2 micrometer (millionth of a meter) to one micrometer in different cells. The rootlet of the stereocilium penetrates the apical surface of the hair cell. The taper at the lower end of the cilium allows it to pivot. The parallel filaments in the core of the stereocilium consist of the protein actin; these filaments are highly cross-linked and make the cilium rigid. The axoneme of the kinocilium, like the axonemes of other true cilia, consists of two central tubules surrounded by nine pairs of microtubules that share a central wall. The axoneme appears most clearly in the horizontal section at the upper right. The vertical section through the kinocilium is through one of the central tubules and the other does not appear. The surface membrane of the hair cell covers both kinds of cilia completely.

of the tympanum is communicated to three small bones in the middle ear: the malleus, the incus and the stapes (the hammer, anvil and stirrup). The stapes, which is the closest to the inner ear, has a pistonlike action that causes the basilar membrane to vibrate. The hair cells are in the organ of Corti, the auditory organ, which lies on the basilar membrane. The hair bundles are inserted into a structure called the tectorial membrane that is placed parallel to the basilar membrane and close to it. The basilar membrane and the tectorial membrane are hinged along different axes; therefore the upward displacement of the membranes caused by the action of the stapes is accompanied by shearing motions between them. These motions displace the hair bundles extending between the two membranes.

As the simplified account presented above indicates, in the functioning of all the sensory organs of the inner ear an initial stimulus (an acceleration or a sound) is transformed into a force on the hair bundle. Since the output of the hair cell is an electrical signal that is sent to the brain along a nerve fiber, the crux of understanding the function of the hair cell is finding out how the displacement of the hair bundle is turned into an electrical response. A complete understanding of the process will undoubtedly require work on the intact inner ear of mammals. Several difficulties, however, stand in the way of relying completely on hair cells from intact mammalian sensory organs. The temporal bone in which the mammalian labyrinth is embedded is extremely hard; mammalian hair cells are particularly sensitive to interruptions of their blood supply; mammalian hair cells are small and difficult to investigate with microelectrodes. Moreover, it is difficult to ascertain the stimulus that reaches the hair bundle and affects it directly when natural stimuli are presented to the ear.

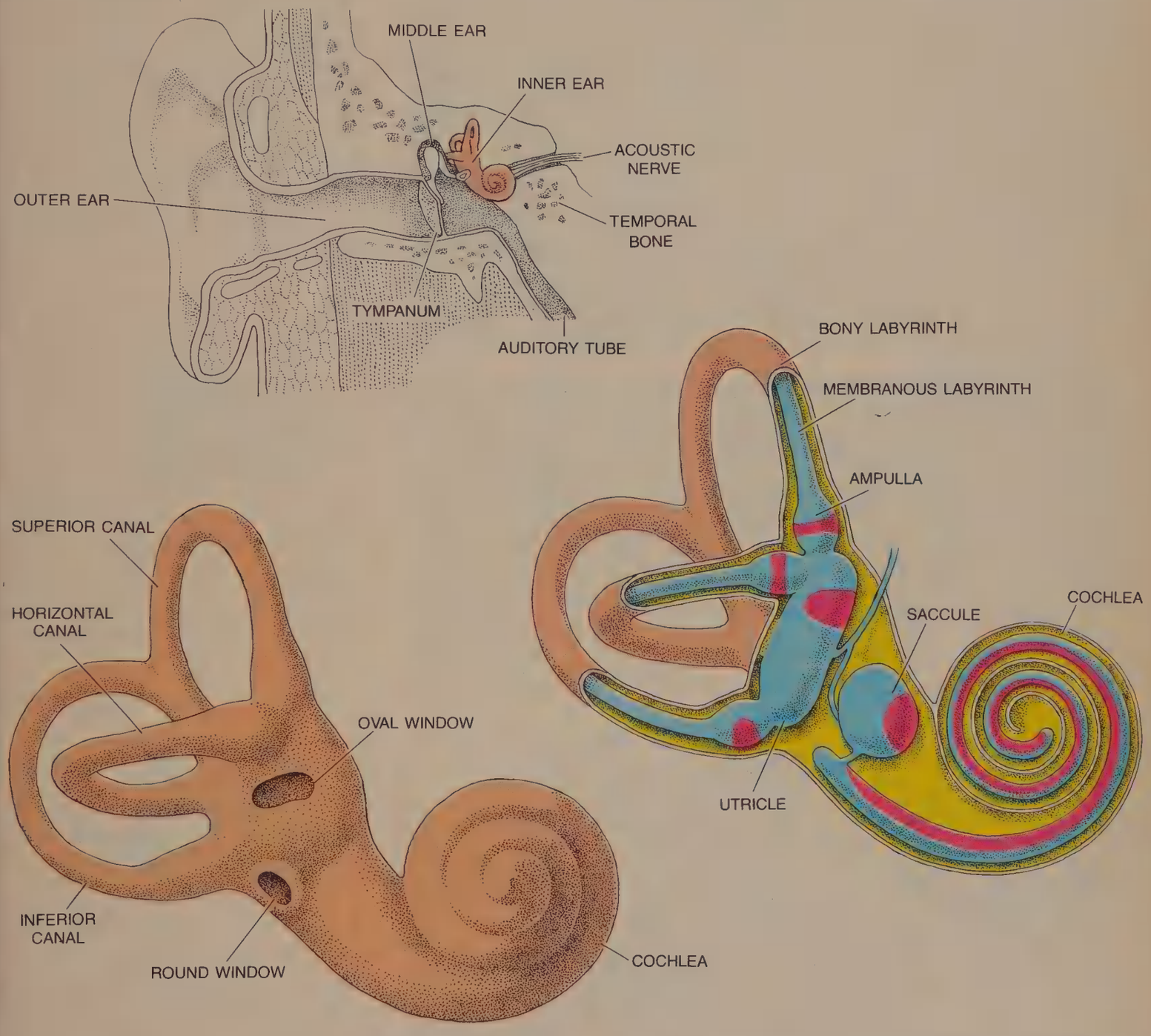
For these reasons preparations of hair cells from lower animals such as fishes, frogs, lizards and turtles have often been exploited in the study of the inner ear. The hardy and relatively large hair cells of the amphibian inner ear make it possible to carry out procedures that cannot yet be done with mammalian hair cells. In much of my own work I have employed hair cells from the inner ear of the American bullfrog (*Rana catesbeiana*). After the saccule is removed from the frog and the otolithic membrane is stripped away the epithelial layer that includes the hair cells is revealed.

The epithelial preparation is put into a physiological solution (a fluid much like the internal fluids of the living organism) and placed under the microscope. A microelectrode is inserted through the apical surface of a single hair cell and into its interior. A micro-

probe is inserted between the kinocilium and the adjacent stereocilia. The probe serves to push the hair bundle back and forth. The motion of the probe can be controlled by a manual "joystick" apparatus or by a computer; in either case the hair bundle can be displaced with high precision. The displacement can also be observed through the microscope. The microelectrode records the electric potential in the interior of the cell and the changes in potential that accompany the movements of the hair bundle.

In order to understand the electrical changes in the hair cell caused by the movements of the hair bundle it is necessary to have some appreciation of the electrical properties of the cell in its unperturbed state. Like the outer membranes of other cells, the plasmalemma of the hair cell is selectively permeable: it allows some molecules to enter the cell and keeps others out. Furthermore, the permeability of the membrane can change through the opening and closing of selective channels. When the channels open, certain ions (atoms or molecules that have lost or gained electrons and therefore have a net electric charge) enter the cell. Through the opening and closing of these channels the cell can regulate its electric potential with respect to the environment.

Molecular pumps in the cell membrane maintain an internal concentration of ions that is different from the concentration in the surrounding fluids. The most intensively investigated of these pumps is the one that reduces the concentration of sodium ions within the cell. Sodium ions have a single unit of positive electric charge. Because the concentration of sodium is lower inside the cell than it is outside, the interior of the cell has a negative potential with re-



HUMAN INNER EAR is the site of six small sensory organs that contain the hair cells. The semicircular canals measure angular acceleration; the utricle and the saccule detect linear acceleration; the cochlea is the organ of hearing. The upper panel is a section through the temporal bone, in which the inner ear is embedded. Because of its shape the inner ear is often referred to as the labyrinth. Sounds striking the tympanum, or eardrum, are communicated to the cochlea by three small bones in the middle ear: the malleus, the incus and the stapes. The panel at the lower left shows the labyrinth as it would appear if it were removed from the temporal bone with a thin layer of bone around it; this is the bony labyrinth. The three doughnut-shaped tubes at the left are the semicircular canals. The spiral structure at the right is the cochlea. The panel at the lower right shows the bony labyrinth opened to reveal the membranous labyrinth, which is composed of the sensory organs. Each organ is a pouch of epithelium. The pouch is filled with the fluid called endolymph (*blue*). Outside the epithelial pouch, between it and the bone, is the fluid called perilymph (*yellow*). The hair cells are found in small patches (*red*). In the semicircular canals the patches are in the region called the ampulla. In the cochlea they are on a membrane that follows the shape of the organ.

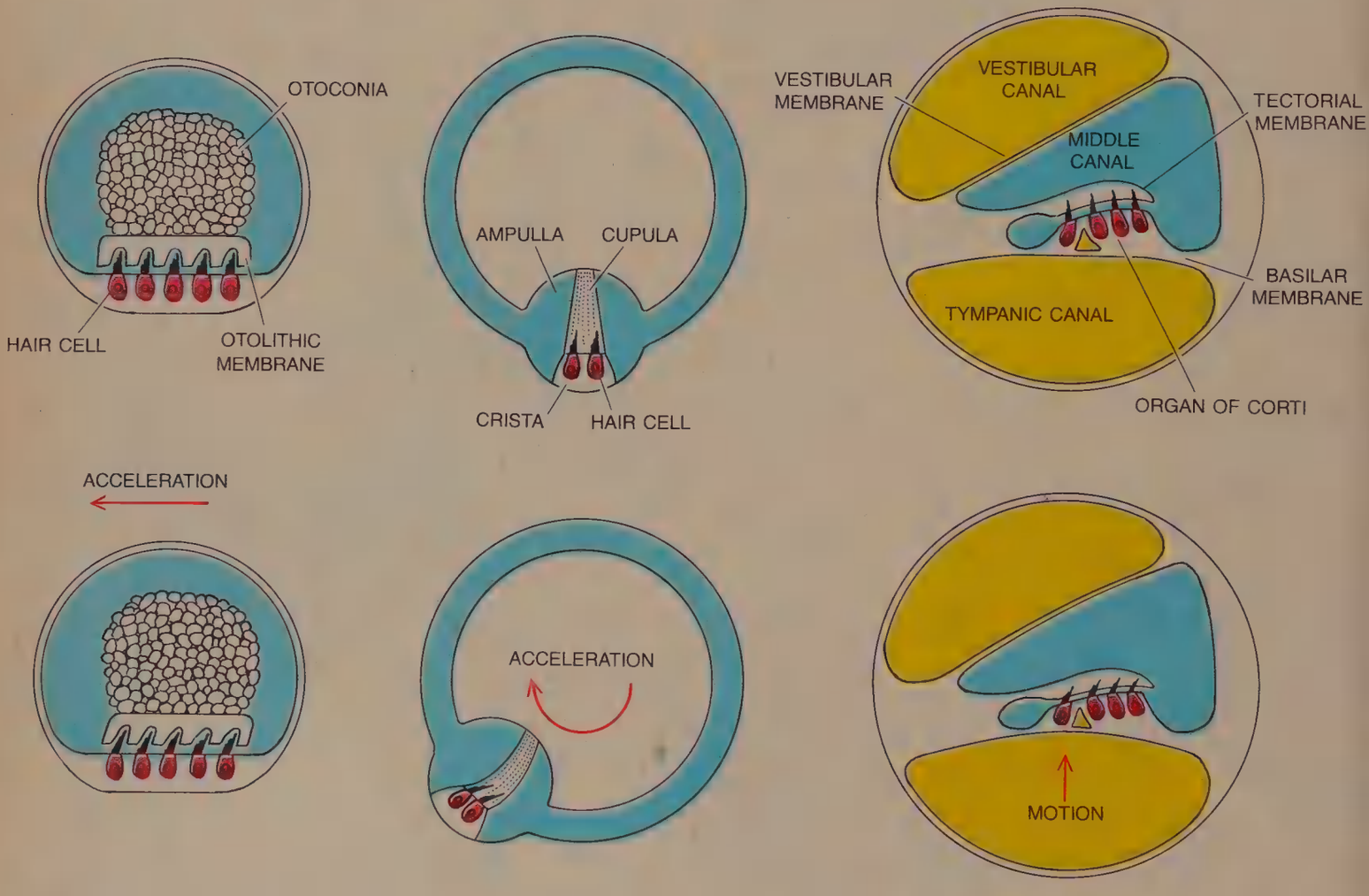
spect to the environment. The difference in potential between the cell and the surrounding fluid is referred to as the membrane potential because it represents the drop in potential across the membrane. It should be noted that if a cell has a negative potential, the interior of the cell attracts positively charged ions. When the channels for such ions open, the ions will rush into the cell, reducing the membrane potential.

The membrane of the cell therefore acts as a capacitor: a thin plate separating electric charges with opposite signs. One property of a capacitor is that it can yield a potential difference by separating a very few charges. Moreover, the passage of only a few charged bodies from one side of the plate to the other can considerably change the potential drop across the capacitor.

The electrical properties of most neurons are regulated by contact with one external fluid. The cell membrane in such cells acts as a single capacitor. In the hair cell the situation is more complex. As we have seen, the hair cell separates the endolymph and the perilymph. The endolymph is in contact with the apical surface of the cell, from which the hair bundle extends. The concentration of positively and negatively charged ions in the endolymph is similar to that in the interior of the cell: high in potassium ions and low in sodium ions. Therefore if the cell were in contact only with endolymph, there would be no potential drop across the membrane.

The basal surface of the cell is in contact with the perilymph. Perilymph has a composition quite different from that of the interior of the cell. In particular, the perilymph has a much higher concentration of sodium ions and a lower concentration of potassium ions than the inside of the cell. If the cell were in contact only with the perilymph, the difference in potential between the interior and the exterior of the cell would be substantial. Indeed, the inside of the cell would have a potential about 90 millivolts lower than that of the surrounding fluid.

Both the apical and the basal surfaces of the hair cell act as capacitors. The membrane potential of the cell as a whole is therefore influenced by contributions from both the apical and the basal areas of the membrane. The contributions are not, however, equal. Both the apical and the basal areas have channels for ions. When the membrane is in the resting, or unperturbed, state, a few channels are always open and there is a small flow of positively charged ions across it. The contribution of the apical and basal areas to the overall membrane potential is determined by the permeability of each area, that is, the number of channels that are open. When the membrane is in the resting state, the permeability of the basal area is substantially greater than that of the apical area. For this reason the membrane potential of the hair cell in the resting state is closer to −90 millivolts than to zero millivolts: the potential is about −60 millivolts.



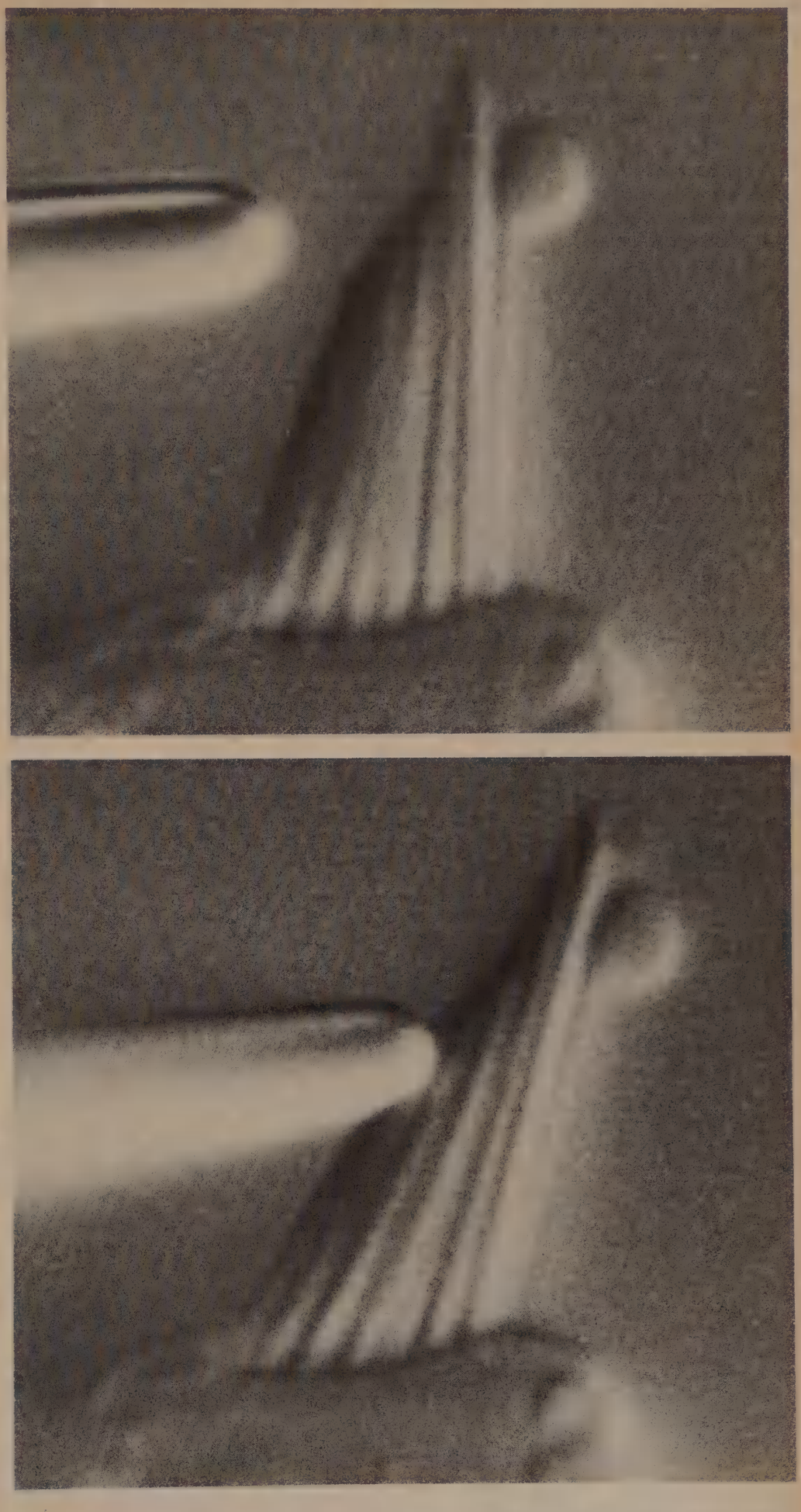
DISPLACEMENT OF THE HAIR BUNDLE is the stimulus that reaches the hair cell in all three types of sensory organ in the inner ear. The three panels of the illustration show how the hair bundle is displaced in the utricle and saccule (*left*), the semicircular canals (*middle*) and the cochlea (*right*). In the utricle and saccule the hair bundles are inserted into a feltwork of protein molecules called the otolithic membrane. Piled on the membrane are the crystals called otoconia. The otoconia are much denser than the endolymph that fills the organ. Therefore when the head is accelerated, the crystals lag behind the fluid. The lagging motion is communicated to the hair bundle by the otolithic membrane. The stimulus applied to the hair cell is thus a displacement of the hair bundle in the direction opposite to that of the acceleration. In the semicircular canals the hair bundles are on a protuberance called the crista and are inserted into a gelatinous sheet, the cupula, that extends across the ampulla. When the head is rotated, the endolymph in the canal lags behind the epithelium of the canal itself. The lagging fluid presses on the cupula and bends it, thereby also bending the hair bundle. In the cochlea the hair cells are on the basilar membrane; the hair bundles are inserted into the tectorial membrane. When the stapes is pulled, both membranes move upward. The membranes are hinged on different axes, so that the movement of the stapes causes a shearing motion that bends the hair bundles.

Therefore when the recording electrode is inserted into a hair cell from the saccule of the bullfrog a measurement of −60 millivolts is obtained. When the hair bundle is pushed with the probe along the axis of bilateral symmetry in the direction of the kinocilium, however, the potential difference decreases to −40 millivolts. The reduction in the potential drop across the membrane is called a depolarization. If the hair bundle is pushed along the same axis in the opposite direction, a quite different change in the membrane potential is observed: it increases from −60 millivolts to −65 millivolts. Such an increase is called a hyperpolarization. If the hair bundle is pushed in either direction along the axis perpendicular to that of bilateral symmetry, there is no change in the membrane potential.

Thus the hair cell is quite responsive to displacements along the axis of bilateral symmetry and is completely unresponsive to displacements along the perpendicular axis. Pushing the hair bundle back and forth along intermediate axes shows that the cell is highly directional in its sensitivity (confirming an inference drawn from extracellular recordings by Otto Lowenstein of the University of Birmingham and Jan Wersäll and Åke Flock of the Karolinska Hospital in Stockholm). If the hair bundle is displaced along intermediate axes, the closer the axis of displacement is to the axis of bilateral symmetry, the larger the cell's response is. Somehow the hair cell is capable of resolving any stimulus into two components: one along the axis of bilateral symmetry and the other along the perpendicular. To the first component the cell responds with a depolarization or a hyperpolarization, depending on the direction of the displacement. To the second component there is no response.

The directional sensitivity of the hair cell is of considerable sensory importance. It implies that whereas a single receptor can provide information about the magnitude of only one component of a stimulus, a group of hair cells with the appropriate sensitivities can provide accurate measurements of an acceleration in any direction. Although the usefulness of the directionality is clear, its mechanism is not known.

Indeed, the phenomenon is a puzzling one in view of the microanatomy of the hair bundle. Individual stereocilia are symmetrical about their long axis; it would appear that bending a cilium in any direction should yield the same result. The kinocilium has a plane of bilateral symmetry, which is the same as the plane of symmetry of the hair bundle. The kinocilium, however, is not responsible for the cell's directional sensitivity. The kinocilium disappears spontaneously from the hair cells of the mamma-



GLASS MICROPROBE is employed to displace the hair bundle in the experimental technique used by the author. The upper micrograph shows a single unperturbed hair cell from the saccule of the American bullfrog (*Rana catesbeiana*). In the lower micrograph the probe has been utilized to displace the hair bundle. The displacement of the hair bundle causes changes in the electrical properties of the hair cell, which in the author's work were measured by means of a small electrode inserted through the upper surface of the cell; the electrode is not shown. These light micrographs were made by differential-interference-contrast microscopy, which exploits differences in refractive index in the parts of the cell to yield bright and dark regions.

lian cochlea in the course of embryonic development. Furthermore, the kinocilium can be removed from the hair cells of the bullfrog's saccule without altering the directionality of the cells' response.

There are three hypotheses that could account for the directional response of the hair cell. The stereocilium could have an asymmetry in its cross section that is not apparent with current techniques of electron microscopy. As an alternative the complex geometric arrangement of the cilia in the hair bundle could somehow confer directionality on an ensemble of structures that do not have such directionality individually. It is also possible that the directional response is a property of the apical surface of the cell and not of the hair bundle.

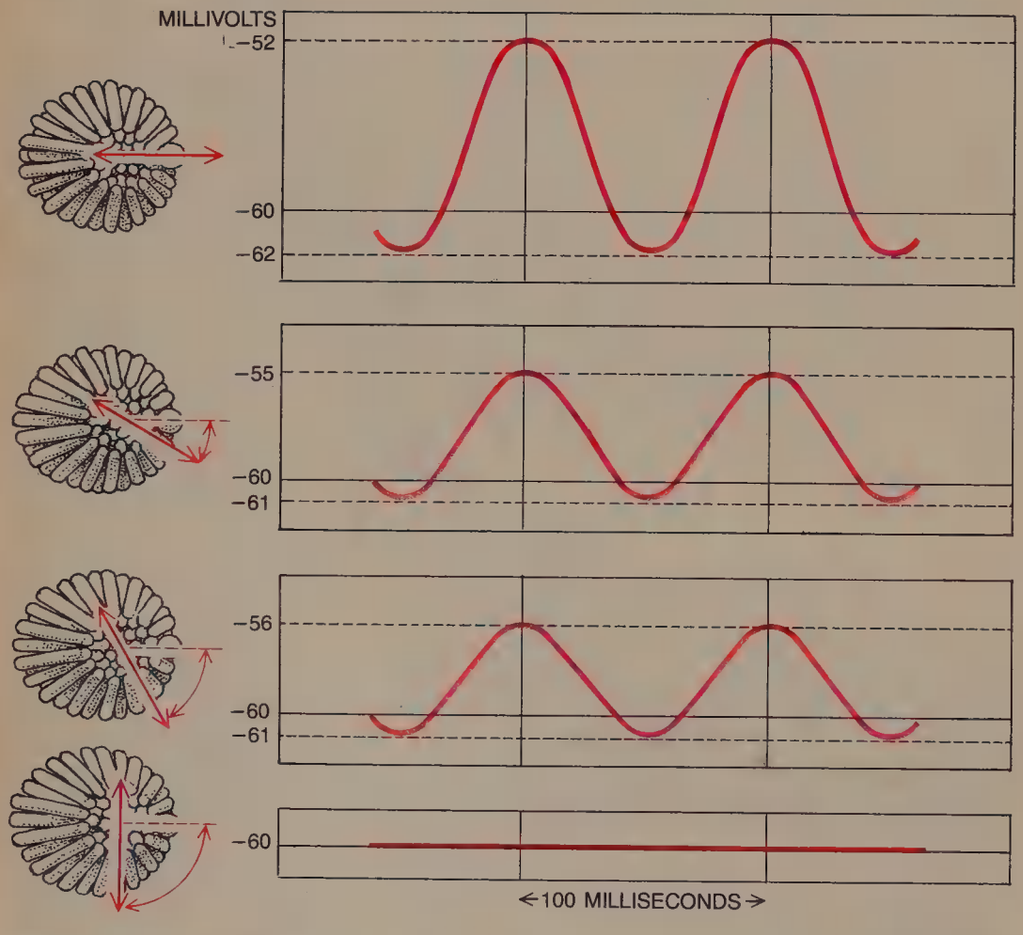
That the anatomical basis of the directionality is not known has not prevented attempts to understand how the complex set of electrical changes is induced in the individual hair cell. As I have noted, the resting membrane potential of −60 millivolts is determined by the ratio of the permeability in the apical area of the membrane to the permeability in the basal area. This suggests that the membrane potential could readily be changed by altering the permeability of the upper or lower area of the membrane. Indeed, both hyperpolarization and depolarization could be accomplished by altering the permeability of one area. Increasing the permeability of that area of the membrane could raise the ratio and decreasing it could lower the ratio, with no change in the permeability of the other area.

My results and those of other workers suggest that in principle this is how the membrane potential of the hair cell is changed. As we have seen, some channels for ions are always open, and more channels are open in the basal area of the membrane than are open in the apical area. When the tip of the hair bundle is moved on the axis of bilateral symmetry toward the kinocilium, additional channels in the apical area open. Positively charged ions, attracted by the negative potential in the cell, rush across the membrane. The increased permeability in the apical area increases its contribution to the overall membrane potential and the potential falls to −40 millivolts. If the hair bundle is pushed on the same axis away from the kinocilium, channels in the apical area of the membrane that are open in the resting state quickly close. The contribution of the apical area to the membrane potential decreases and the potential difference increases to −65 millivolts.

The "transduction channels" in the apical area of the membrane that are responsible for the change in potential do not appear to be highly selective. When the channels are open, they allow most positively charged substances .6 nanometer (billionth of a meter) or less across to enter the cell. The main positively charged ion in the endolymph, however, is potassium, and it seems that potassium is the main carrier of electric charge across the membrane. Experiments in which the potassium in the endolymph was replaced with other positively charged ions show that although other substances can carry the current, potassium probably carries most of it in the intact hair cell.

Since the molecules that enable current to flow across the membrane have not been conclusively identified, it is difficult to specify the location of the transduction channels with precision. There is, however, some indirect evidence for their location. When the transduction channels open, small electric currents flow though the fluid around the cell as well as across the membrane. By means of a recording electrode the current at various points around the perimeter of the hair bundle can be measured as the cilia are being moved. The current in the surrounding fluid is greatest at the top of the hair bundle, near the tips of the stereocilia. This suggests that when the hair bundle is deflected, channels open not at the base of the cilia where they pivot but near their tips, perhaps where adjacent cilia touch each other. Such evidence is far from conclusive and additional work is needed before the hypothesis can be accepted.

Electrical recordings from hair cells show that the initial response of the receptor is extremely fast. The fact that the auditory system of vertebrates such as bats and whales can detect tones with frequencies as high as 100 kilohertz implies the transduction apparatus can also respond very fast. Recordings from isolated inner ears confirm this implication. The hair cells of the bullfrog's ear normally respond to sounds with frequencies no higher than three kilohertz, but these cells begin to respond within a few tens of microseconds (millionths of a second) after the stimulus is applied. Hence the opening of the transduction channels, which initiates the cell's response, takes roughly the same time as a molecular event such as the action of an



DIRECTIONAL SENSITIVITY of the hair cell is shown in traces representing recordings made with an electrode inserted into a single receptor from the saccule of the bullfrog while its hair bundle was being moved. The figures are based on experimental records made by Sandra L. Shotwell in the author's laboratory. When the hair cell is in its resting state, its interior has an electric potential 60 millivolts lower than that of the surrounding fluid. If the hair bundle is displaced toward the kinocilium along the axis of bilateral symmetry, the potential difference decreases to −52 millivolts (*top*). If the bundle is displaced away from the kinocilium along the same axis, the difference increases to −62 millivolts. As the direction of displacement diverges from the axis of symmetry the size of the response decreases (*middle panels*). The cell does not respond to displacements along the axis perpendicular to that of symmetry (*bottom*).

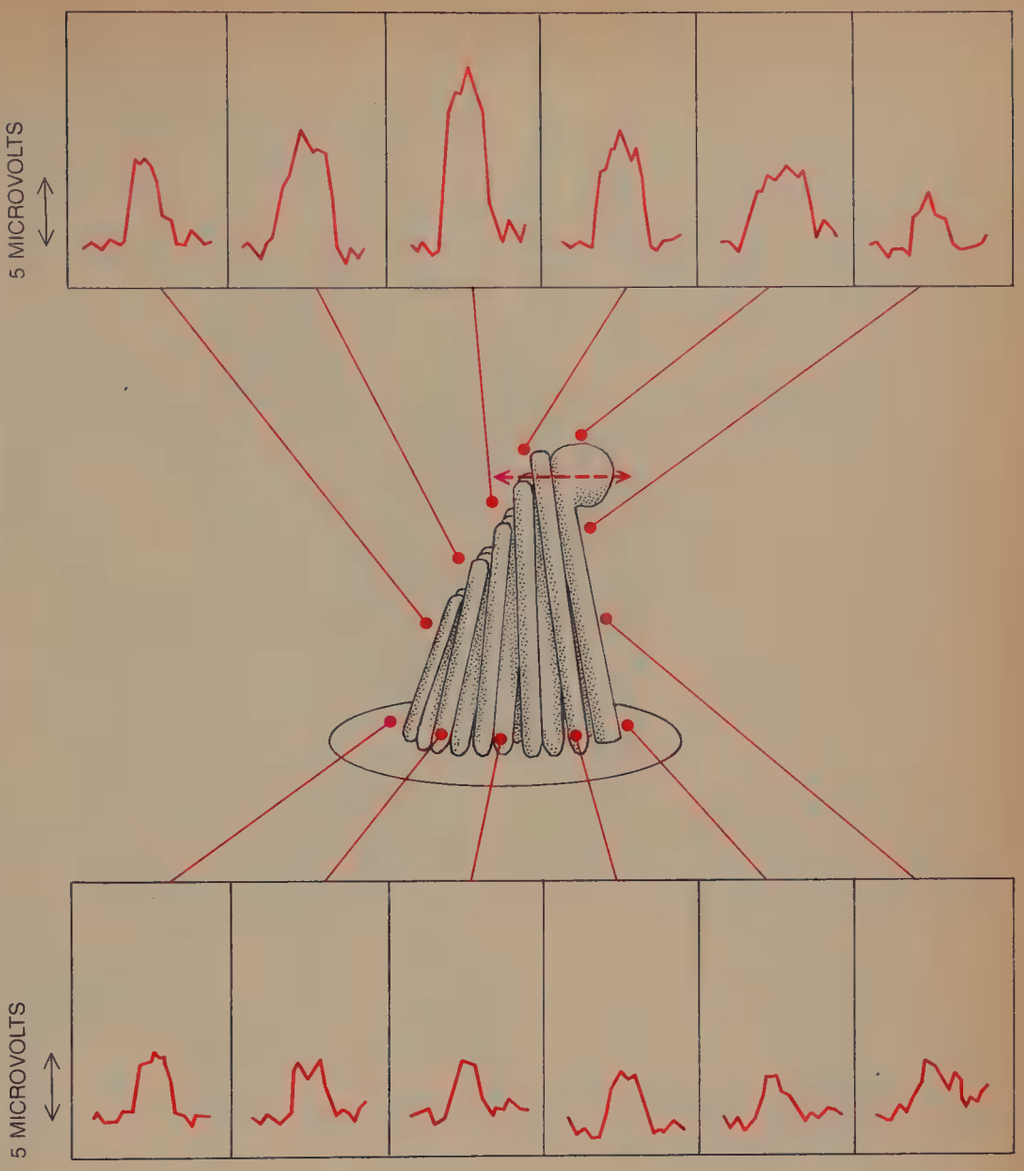
enzyme. It is unlikely that many processes intervene between the application of the force to the hair bundle and the initiation of the cell's response; it is as if the mechanical force were acting directly on the transduction channel.

A very small displacement of the hair bundle is sufficient to cause the hair cell to respond. The hair cells of lizards and turtles begin to respond when the bundle is moved a distance estimated to be as little as 340 picometers. In the mammalian cochlea the estimated threshold is even lower: about 100 picometers. Recordings of hair cells from the saccule of the bullfrog show that the entire range of the cell's response comes over a deflection of about .5 micrometer. Since the hair bundle in the bullfrog's saccule is about 8.2 micrometers long, the entire response comes over an angular deflection of only three degrees.

It seems likely that a large continuous stimulus (such as one that would result from a rapid acceleration) could deflect the stereocilia much more than three degrees. The cell might then have a high constant output. Subsequent small changes in the stimulus from the high initial level would not change the cell's response. Such a pattern would be disadvantageous to the organism because the small changes are of potentially greater significance than the large constant stimulus is.

In the saccule of the bullfrog, however, such sensory "saturation" does not occur. When an isolated hair cell from the saccule is tested in the unperturbed state, it is found to be most sensitive to displacements of a few tenths of a micrometer in either direction. If the hair bundle is moved to a new rest position that is, say, one micrometer from the first position, the result is not simply a large constant response. Instead after a few tenths of a second the response to the static one-micrometer displacement decreases and the cell becomes sensitive to small movements away from the new rest position. The basis of this adaptation is not known, but its significance is clear: it enables the organism to disregard large continuous stimuli while maintaining great sensitivity to small, rapid sensory inputs that are of greater importance.

The responses of the isolated hair cell induce wonder at a molecular mechanism capable of reacting to stimuli with an amplitude of picometers that come as often as 100,000 times per second. The means by which the transduction channels are opened and closed is currently the subject of intense scrutiny in my laboratory at the California Institute of Technology and in other laboratories. One plausible hypothesis is that the transduction channels are continually being opened and closed by the random



PATTERN OF ELECTRIC CURRENTS in the fluid around the hair cell suggests that the channels mediating the cell's electrical response are near the top of the hair bundle. The changes in electric potential in the hair cell are caused by flows of positive ions into the cell from the surrounding endolymph; the ions are attracted by the negative potential in the interior of the cell. The main positive ion is potassium. The flow of potassium ions (and hence the electric potential of the receptor) is regulated by the opening and closing of "transduction channels" in the cell membrane. When the transduction channels open and potassium ions flow into the cell, there is also a small current in the endolymph around the channels. Measurements of the voltage due to this current at various points around the hair bundle are shown. The largest of the currents has an amplitude of about 13 microvolts (millionths of a volt). The fact that the current is greatest near the top of the hair bundle suggests the transduction channels are there.

thermal motions of nearby molecules. If that is the case, the fraction of channels open at any one moment will depend on the difference in energy level between a closed channel and an open one. If the difference is greater than the kinetic energy of the average collision with a nearby molecule, the opening of a channel will be a rare event and at any moment most channels will be closed. If the energy levels of closed and open channels are similar, however, about half the channels will be open at any given time.

Suppose that initially the energy level of an open channel in the apical area of the hair-cell membrane is relatively large. Most of the channels will then be closed. If the displacement of the hair bundle somehow reduces the energy of an open channel, the fraction of channels that are open will increase rapidly under the influence of the motion of the nearby molecules. David Corey of the Yale University School of Medicine, who was then a graduate student in my laboratory, made recordings from hair cells that are consistent with this hypothesis. The central question of how the movement of the hair bundle changes the energy of the channel, however, remains unanswered.

After considering what is known about the transduction mechanism of the hair cell, the question naturally arises of how the change in potential leads to the conveying of a message to the brain. Since hair cells do not have axons, they cannot relay information directly. Instead the message is carried by a neuron

with an ending near the basal surface of the receptor. When the transduction channels of the hair cell open, the potential drop across the membrane changes very fast in all parts of the cell. In the basal area of the cell membrane there are many channels that selectively admit calcium ions. The action of the calcium channels is voltage-dependent: when depolarization occurs, the channels open.

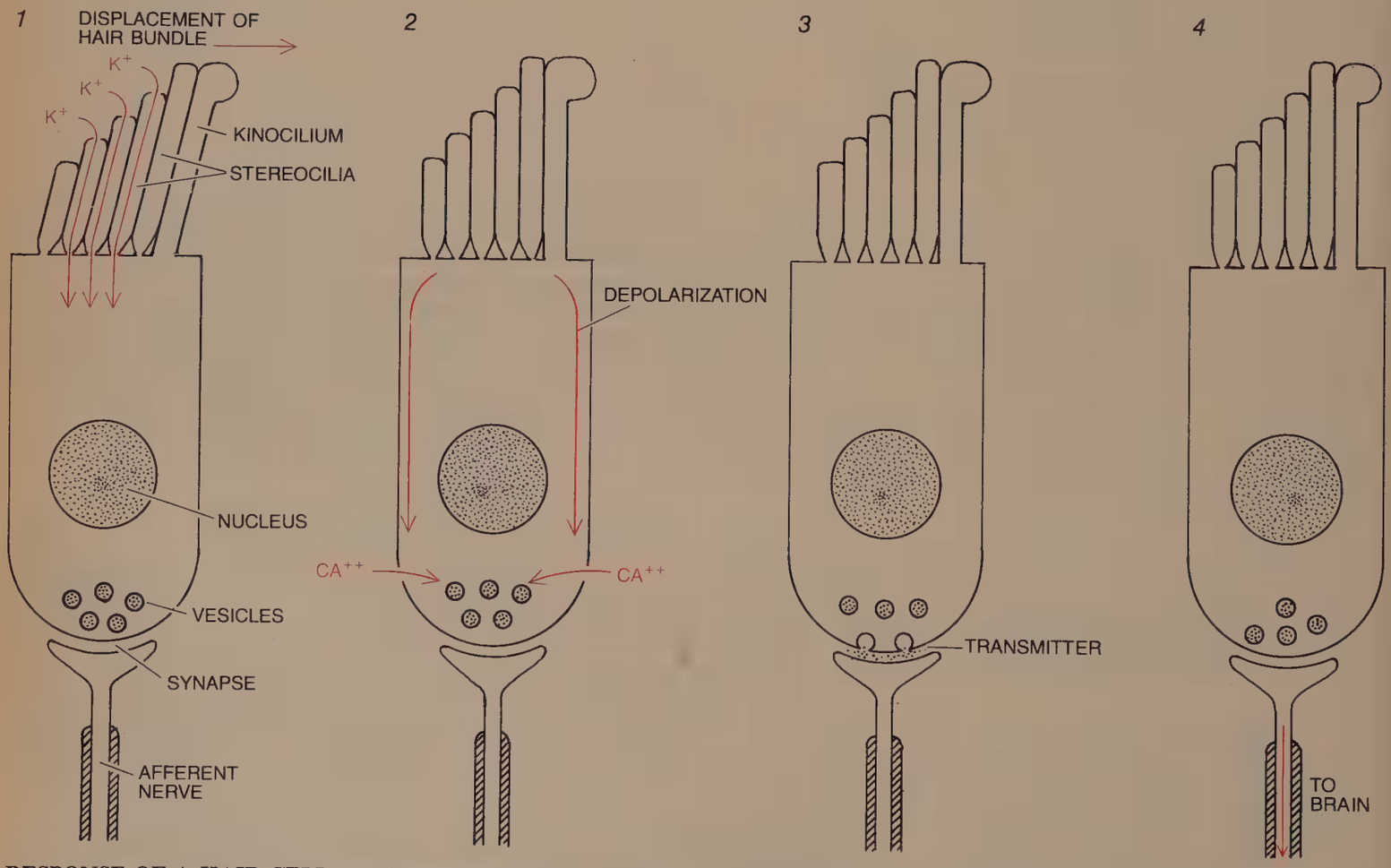
Once calcium ions are admitted to the cell, they serve to trigger another significant event. Near the base of the hair cell are many vesicles containing a chemical substance that acts as a neurotransmitter. The calcium ions cause the vesicles to fuse with the outer membrane and release their contents into the synaptic gap that separates the hair cell from the terminal of the neuron.

The exact chemical composition of the transmitter substance is not known. It is known, however, that when the transmitter molecule is released from the vesicle, it diffuses across the synaptic gap: a distance of about 20 nanometers. On combining with receptors on the membrane of the nerve terminal the transmitter causes a depolarization of the neuron much like the depolarization of the hair cell. In the neuron the depolarization elicits one or more "action potentials," electrical signals that travel along the axon.

A hyperpolarization of the hair-cell membrane has the opposite effect. Even in the resting state some of the calcium channels at the base of the cell are open; hence a small amount of the transmitter is always diffusing across the synapse and a small signal is carried to the brain. A hyperpolarization closes down the open calcium channels and reduces the amount of transmitter that is released. The frequency of occurrence of the action potentials in the neuron is thus reduced, and in the logic of the nervous system such a reduction can also serve as an information-carrying signal.

The information from all the hair cells in the inner ear is carried to the brain by axons in the eighth cranial nerve. Each of the many fibers in the nerve carries signals from hair cells of only one organ. Moreover, each fiber carries very specific information. For example, a particular fiber from one semicircular canal is excited only by angular accelerations about a particular axis in the head, and only by the accelerations in the clockwise direction.

Messages travel not only from the hair cell to the brain but also from the brain to the hair cell. Near the base of the hair cell are two sets of nerve fibers. The fibers called afferent carry information to the brain; those called efferent carry information back to the hair cell. When the efferent fibers to the cochlea are activated, the hair cells appear to become less sensitive to sound. In addition they seem to be less sharply tuned to particular frequencies. The significance of such desensitization is not clear. It is one of the many puzzles of the hair cell that remain to be solved. Work is being done on hair cells in many laboratories, and in the next few years many questions, including the central one of the transduction mechanism, will undoubtedly be answered. Further work, however, can only reinforce a sense of wonder at the sensitivity and complexity of this miniaturized piece of biological apparatus.



RESPONSE OF A HAIR CELL culminates in the transmission of an electrical signal to the brain along the afferent nerve fiber at the base of the cell. The response is shown schematically in the panels of this illustration. When the hair bundle is displaced, the transduction channels open (*1*). Potassium ions flow into the cell, and the potential difference between the cell and the surrounding fluid falls. The reduction in potential difference, which is called a depolarization, spreads almost instantaneously through the cell (*2*). In the lower part of the cell are channels that selectively admit calcium ions. The depolarization causes these channels to open, whereupon calcium flows in. Near the base of the cell are vesicles containing a neurotransmitter: a substance that transmits the nerve impulse from one cell to the next. The calcium ions cause the vesicles to fuse with the basal part of the hair cell's surface membrane (*3*). In fusing the vesicles release the neurotransmitter they contain. The transmitter, whose chemical nature is not known, diffuses across the synaptic space between the hair cell and the neuron; it then excites the neuron, which sends a message to the brain along a fiber of the eighth cranial nerve (*4*).

We discovered a new kind of brain cell.

You'll find it at places like the Princeton Advanced Center for Learning, where experts from a dozen different disciplines cross-pollinate their thoughts and ideas.

But as far as companies in America go, we at Kodak may be the only ones who can bring together our own experts in programming, optics, physics, organic and polymer chemistry, fibers, human engineering, imaging, marketing, paper technology ...we've probably skipped a few.

Call it synergism or whatever. But when these pros sit down at our table, they come up with some truly incredible results.

The Kodak Ektaprint 250 duplicator is a good case in point. It has grown out of more disciplines than any single Kodak product before it.

The nucleus is two 8085 microprocessors, with a combined yield of over 100,000 bytes, and responsibility for 134 separate functions.

There are all kinds of thoughtful touches. One is a programmable display that's as friendly as can be—with 94 different messages, not codes, to help anticipate and satisfy operator needs. And there's an internal diagnostic system that's amazingly specific.

Then there's single-pass duplexing. Lots of heads went together for this run-through-once, print-both-sides convenience. Collating and cover inserting, too.

A couple of other neat feats: Fiber specialists from our Eastman Chemicals Division in Tennessee helped create the "perfect" toner brush. In Rochester, our experts in 35 mm added sprockets to the belt drive. And a reflective "bathtub" was perfected by our processing group; it helps keep the paper path shorter and virtually trouble-free.

With all these new capabilities, you may think the 250 needs a genius to run it. Wrong. And that's the beauty. 99% of the duplicator is microprocessor-controlled. That means reliability, flexibility, productivity, easier operation. And time to think about more important matters.

For more information on how we plan to combine our talents for future applications in microelectronics, please send for our brochure ..."Resources To Innovate, Commitment To Serve..." (B3-190). Write to Eastman Kodak Company, Dept. GB-SA 6, 343 State Street, Rochester, New York 14650.

Kodak

Kodak. Where technology anticipates need.

© Eastman Kodak Company, 1983

SCIENCE AND THE CITIZEN

Orphans in the Marketplace

In a free-market economy if the demand for a commodity is not sufficient to ensure that its sale will yield a profit, the commodity will not be produced. If the commodity is a medication and the demand for it is the result of a disease, the consequences of the economic system can be life-threatening. The rule of thumb used by American pharmaceutical manufacturers is that there must be 100,000 victims of a disease to make the distribution of a drug for it profitable. There are about 2,000 diseases that afflict fewer than 100,000 people in the U.S. For the most part the pharmaceuticals industry has shown little interest in developing treatments for them. For some of the diseases medications exist but are not currently on the market.

Such orphan diseases and orphan drugs were the subject of a recent conference at the University of Michigan at Ann Arbor. It was held at a time when there is considerable interest in the subject. Bills have recently been adopted by the House of Representatives and the Senate providing tax incentives for companies that bring orphan medications to market. The U.S. Food and Drug Administration (FDA) has established an Office of Orphan Product Development to make it easier for such substances to be approved for clinical use without the support of a sponsoring corporation.

Some observers maintain that the stringent and complex FDA regulations governing premarket drug testing are responsible for the orphan status of many therapeutic substances. The regulations are said to make the cost of developing a new drug so high that it inhibits the testing of new medications for rare conditions. George J. Brewer of the University of Michigan Medical School, who was one of the organizers of the conference, disagrees with this assessment. According to Brewer, "overregulation is not the problem with orphan drugs. ... The FDA is willing to streamline the approval procedure."

Brewer argues that the main causes of the orphaning of drugs and diseases are an absence of interest on the part of the pharmaceutical companies and a lack of work on the diseases by biomedical investigators. As evidence of the manufacturers' lack of interest he has cited three substances that show promise for the cure or control of orphan diseases. One substance is L-5-hydroxytryptamine, an experimental drug for the treatment of myoclonus, a rare neuromuscular disorder. Another is cysteamine, which is employed in treating cystinosis, an inherited defect of amino acid metabolism that affects about one infant in 50,000.

The third medication is zinc acetate, a compound Brewer is testing as a means of controlling the symptoms of sickle-cell disease. Sickle-cell anemia, the result of an inherited defect in red blood cells, affects about 25,000 black Americans. Brewer's work suggests that zinc compounds can control one of the major symptoms of the disease, an abnormally high level of calcium in the red blood cells. No pharmaceutical company has shown much interest in any of the three substances, Brewer said, even though the drugs have considerable therapeutic promise.

In the case of zinc acetate the obstacles to a profitable investment are increased by the fact that the active agent, zinc, is a chemical element and therefore cannot be patented. If a company took the steps necessary to have the compound approved by the FDA for therapeutic use, the company could not claim exclusive rights to manufacture the medication. Indeed, a second company, having been spared the cost of developing the product, might be able to sell it at a lower price.

The number of drugs that offer an effective means of curing or controlling a disease but are not available commercially is a matter of controversy. Pharmaceutical companies are under no obligation to reveal how many substances they have tested but not placed on the market. At the Ann Arbor conference representatives of the Pharmaceutical Manufacturers' Association said that only about 10 orphan drugs had been found in a survey of manufacturers and biomedical investigators. According to Brewer, however, the survey was incomplete. Neither he nor other investigators he knows who have worked on orphan drugs were approached, he said. There may be considerably more than 10 orphan products, he added.

Furthermore, cures or effective therapies are known for only about 10 percent of the 2,000 orphan diseases. Workers in the biological sciences have shown little enthusiasm for identifying and testing potential medications for the remaining 90 percent. One reason is the scarcity of Federal funding for the investigation of medications and specific treatments. Most public funding in biomedicine has gone for work on disease mechanisms or on basic physiological processes. The development of drugs and treatment regimens has been left to the manufacturers.

Brewer argues that if treatments for orphan conditions are to be developed, the National Institutes of Health, the main Federal agency supporting work in the biological sciences, will have to establish a separate review process for making grants in this area. Grant applications for work on specific drugs and treatments suffer by comparison with applications for work in basic biological processes. "Our grants get thrown in with cutting-edge science, which is quite elegant," he said. Work on varying the dose of a medication for treating an orphan disease can appear "quite mundane" in such a context.

Brewer added that without increased public support there may never be effective medications for most orphan diseases. Pharmaceutical manufacturers cannot fundamentally alter their policy toward the diseases because the companies' continued existence depends on their capacity to make a profit. According to Brewer, "the drug companies can be helpful in the final stage [of drug testing and marketing], but biomedical investigators must do much of the first stage...."

Element 109 and Beyond

The reported synthesis of a new heavy element—number 109 in the periodic table of the elements—has stirred fresh interest among nuclear physicists in the prospects for finding a long-sought sequence of stable, superheavy elements with atomic numbers higher than 114. A single atom of element 109 was tentatively identified among the products of an extraordinarily "gentle" nuclear collision staged last summer at the laboratory of the Society for Heavy Ion Research in Darmstadt, West Germany. The team of experimenters was led by Peter Armbruster and Gottfried Münzenberg.

The new element was produced in the laboratory's linear heavy-ion accelerator (named Unilac) by a technique known as cold fusion. A beam of iron-58 ions was accelerated and fired at a thin foil of bismuth 209. In principle a collision between an iron nucleus with 58 nucleons (26 protons and 32 neutrons) and a bismuth nucleus with 209 nucleons (83 protons and 126 neutrons) could be expected to result in the formation of a compound nucleus with 267 nucleons (109 protons and 158 neutrons), but only under certain strictly defined conditions. The collision would have to be energetic enough for the two nuclei to fuse but not so violent that they would immediately fly apart.

Calculations showed that even under carefully controlled conditions the cold fusion of iron 58 with bismuth 209 would take place only rarely: about once in every 100 trillion nuclear collisions. Armbruster, Münzenberg and

THE BMW 733i. IT DOESN'T SCREAM WEALTH. IT JUST QUIETLY REWARDS INSIGHT.

It's no secret that there are some expensive sedans that are purchased for their sheer expensiveness.

Sedans that serve as a sort of public declaration of prosperity and status.

Such cars are the antithesis of the BMW 733i—the luxury sedan engineered in the belief that performance, not self-congratulation, is the ultimate measure of a car's true worth.

A PANACEA FOR PURISTS.

The 733i is a car whose subtleties of engineering and technology place it beyond the needs of all but a handful of purists. (Which is perhaps just as well, given its limited supply.)

For those serious drivers, however, it is as close to a necessity as a $35,000* sedan can come.

Its fuel-injected engine delivers the sort of heady response that suggests an oversized power-plant beneath the hood.

Not so. The 733i's 6-cylinder engine arrives at high performance through refinement, not brute force.

A system of microprocessors gathers data on fuel mixtures, engine speed and other factors. It then orders ignition to occur at the optimum moment—accomplishing through elegance of design what might otherwise require additional cylinders and extra girth.

Its suspension handles the landscape with precision instead of bluntness—removing almost surgically the wallowing, skittishness and related uncertainties from bumps and curves.

(One reason: a new rear suspension that incorporates one of the most significant breakthroughs in independent rear suspension design in a generation.)

The 733i is also designed for those who prefer the subtleties of road topography to the anesthetizing ride of limousine-like sedans. Its highly tactile steering places you in touch with that alternately pleasant and harsh reality called the road—filtering out just enough of its harshness, while never removing the reality.

The result is a machine that is as much system as car, one so perfectly balanced and rigorously self-policing that it even calculates when routine service is needed.

ENGINEERED FOR DRIVING, NOT JUST SITTING.

Inside, the 733i is engineered for activities of considerably more subtlety than sitting.

Its interior literally helps you drive, rather than merely accompanying you on the trip.

Its supple leather seats are designed to follow the curvature of the spine, placing the driver in postures that are anatomically correct and help reduce driver fatigue.

An Active Check/Control provides readings on 7 different measures of the car's operational readiness.

An on-board computer provides all manner of trip information, such as the distance remaining to your destination.

In fact, no detail in a BMW 733i ever suffers from inattention because it's judged 'minor.' And its warranty reflects this obsessiveness. The 733i is backed by BMW's 3-year/36,000-mile limited warranty and a 6-year limited warranty against rust perforation.†

All of which led Car and Driver to conclude that the 733i's "parts and pieces...work so well together they must have been melded in another world."

It may be seen, however, in much more convenient locations. We invite you to visit your nearby BMW dealer, who will be happy to arrange a test drive.

THE ULTIMATE DRIVING MACHINE.

*Manufacturer's suggested retail price: $34,300. Actual price depends on dealer. Price excludes state and local taxes, dealer prep, destination and handling charges. †Warranty applies to automobiles purchased from authorized U.S. BMW dealers only. See your BMW dealer for details. © 1982 BMW of North America, Inc. The BMW trademark and logo are registered. European Tourist Delivery can be arranged through your authorized U.S. BMW dealer.

their colleagues planned a 10-day run on the Darmstadt accelerator in late August to search for evidence of element 109. It was not until August 29, the last day of the scheduled run, that their detector registered a likely event. Four measurements of the properties of the newly formed atom and the products of its decay confirmed that a nucleus of element 109 had evidently existed for approximately five milliseconds.

The synthesis of element 109, assuming that it can be repeated, would set a new nuclear record: it is the heaviest element detected so far. The natural elements extend through atomic number 92, and artificial ones through number 106 have been known for some time. In 1981 element 107 was identified by Armbruster and Münzenberg and their colleagues, following an earlier, disputed claim by investigators at the Joint Institute for Nuclear Research in Dubna. (Element 108 is thought to be less stable than either 107 or 109.)

The apparent success of the cold-fusion technique has stimulated efforts to extend it to the long-standing search for superheavy elements. The possibility that such elements might exist is suggested by the shell model of the nucleus, which describes nucleons as occupying successive shells much like the shells of electrons that determine the chemical properties of atoms. Each shell can hold a particular number of nucleons, and filled shells bring enhanced stability. The filling of nucleon shells in elements with atomic numbers in the range from 114 to 126 is expected to make such hypothetical nuclei unusually stable. Elements occupying such an "island of stability" would presumably be easier to detect than elements with atomic numbers ranging from 109 to 114.

The German physicists have now joined forces with a group at the Lawrence Berkeley Laboratory, headed by veteran heavy-element synthesizer Albert Ghiorso, in an attempt to make element 116 by the cold-fusion technique. The current experiments call for bombarding a target of curium 248 with an accelerated beam of calcium-48 ions in the Berkeley laboratory's SuperHILAC heavy-ion accelerator as well as in the Darmstadt machine. Ghiorso, who once bet his colleague Glenn Seaborg $100 that superheavy elements would never be found, has reportedly lowered the odds against the experiments' succeeding to no more than 100 to one.

Rosy Future

To understand the genetic control of development, molecular biologists are eager to find ways to introduce a gene into animal germ cells and monitor its expression as a protein product in the mature animal. Ideally the inserted gene would become a permanent part of the animal's genome, or total complement of genes; it would be passed on to successive generations and be expressed by the animal's offspring. Recombinant-DNA techniques make it possible to isolate a single gene, perhaps to alter it in a defined way and to clone it in quantity in bacteria. The trick then is to get the gene into animal germ cells and have it integrated into the genome without further, inadvertent alteration, so that it is correctly expressed, stably inherited and expressed in future generations. This has now been accomplished. Workers in the Department of Embryology of the Carnegie Institution of Washington report in *Science* an efficient new method of gene transfer whereby a cloned gene can be injected into embryos of the fruit fly *Drosophila melanogaster*. When the gene called *rosy* is transferred into embryos of a mutant line of flies with brown eyes, fertile adults that develop from the embryos give rise to a stably modified strain with the bright red eyes of normal *D. melanogaster* flies.

In earlier (and continuing) efforts to insert foreign DNA into animal cells investigators have "transfected" cultured cells by exposing them to naked foreign DNA, they have microinjected the DNA into the nucleus of cells or into fertilized eggs and they have exploited the ability of an animal virus to "transduce" a DNA segment into cultured cells. Gerald M. Rubin and Allan C. Spradling of the Carnegie Institution took advantage of a different natural vector: a transposon, or transposable genetic element. A transposon is a segment of DNA that moves around in the genome of a bacterium or a higher organism and is capable of inserting itself—and often other segments of DNA incorporated within it—at a large number of sites.

Rubin and Spradling worked with a family of *Drosophila* transposons called *P* elements. Rubin and several co-workers had identified a particular *P* element, much longer than the rest of them, that seems to be required for transposition. It is long enough to code for a transposase: an enzyme that catalyzes the transposition of the long element itself and of shorter "defective" *P* elements. The long element also encodes a repressor protein that blocks transposase activity. The result is that the introduction of a long *P* element into germ cells not already carrying it gives rise to transposition at a high rate; once a long element is present the repressor accumulates and prevents transposition.

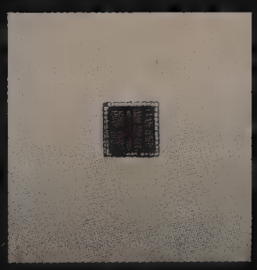
In a first series of experiments Spradling and Rubin inserted a long *P* element into a plasmid (a small circle of bacterial DNA) and injected the plasmid into a *Drosophila* embryo. They were able to show that the *P* element can catalyze its own transposition into the embryo's germ cells. Plasmids are routinely used as vectors for introducing foreign DNA into bacteria. The self-transposing ability of the long *P* element suggested that a gene of interest might be introduced into *Drosophila* germ cells as a passenger within a long *P* element. Better yet, unwanted effects of the transposase could be avoided by inserting the gene into a defective *P* element and supplying just enough transposase to catalyze the desired transposition.

Rubin and Spradling undertook to transpose the *rosy* gene, which codes for the enzyme that makes a normal fly's eyes red. They inserted *rosy* into a short *P* element, which was inserted in turn into a plasmid and was cloned in bacteria. Then they microinjected the plasmid into embryos of mutant (brown-eyed) strains. To supply the necessary transposase they tried two stratagems. One was to inject the plasmid carrying the *rosy* transposon into mutant embryos derived from a cross between males carrying the long *P* elements and females lacking them. The genome of such a hybrid undergoes transposition at a high rate; the transposition is catalyzed by the male parent's transposase, which is temporarily "derepressed" in the hybrid germ cell. The other stratagem was to coinject, into embryos lacking *P* elements, a high concentration of the *rosy* plasmid and a low concentration of the long *P* elements, which would supply the transposase. Given the excess of *rosy* transposons and the limited supply of transposase, the expectation was that many germ cells could incorporate a *rosy* transposon but not a long *P* element.

Both methods worked. Some 8 percent of the injected embryos developed into fertile adults, and 39 percent of those adults had at least some progeny with red eyes; the eye color was inherited stably in later generations. By probing the DNA of the red-eyed flies Rubin and Spradling were able to show that the *rosy* transposon was indeed integrated into chromosomal DNA and was responsible for the change in eye color. Furthermore, only the transposon was taken up by the fly genome; the bacterial DNA of the plasmid was not.

Rubin and Spradling foresee two major applications of their technique. One is the study of gene regulation in higher organisms, notably in *D. melanogaster*. The fruit fly has been the subject of intensive genetic analysis for decades. Many genes that have been identified seem to play an important role in development, but their protein product and mode of operation are not known. Now those genes and their products can be studied in detail in the developing organism. The fly's *P* element or modified versions of it may function in other animals as well; if the *P* element will not cross species lines, analogous transposons should be found in other species. Eventually, Rubin and Spradling think,

In 1981, Hewlett-Packard announced the world's densest computer chip.



Our 450,000 transistor, one chip
32-bit CPU.

Today, it's the heart of a 32-bit that's so affordable your top engin can have their own mainframes.

From time to time, miracles of technology come along to make previously impossible tasks not only possible, but easy. That little integrated circuit chip on the preceding page is one of those technological miracles.

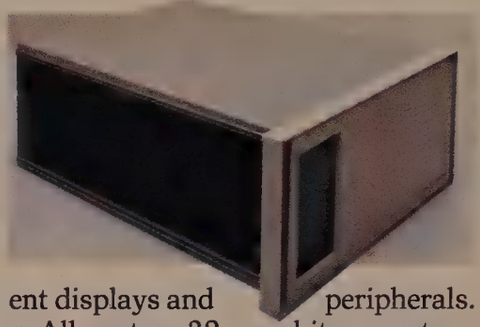
Hewlett-Packard didn't develop it just to break the record for most transistors on a chip, but to put on an engineer's or scientist's desk a computer so powerful that it can do the work of mainframes costing four times as much.

32-bit computers for 32-bit applications.

The new HP 9000 computer based on this and four other 'superchips' can handle formidable engineering and scientific problems. The scientist solving complex systems of equations, the mechanical engineer doing finite element analysis or three-dimensional modeling, the electrical engineer analyzing complex circuits or designing very large-scale integrated circuits — these are the kinds of technical people and problems the HP 9000 family is designed for.

It comes in three versions. The integrated workstation is complete with keyboard, color or monochromatic graphics display, fixed and flexible disc drives, and printer. For systems manufacturers, there's a rack-mountable box. And for a variety of single-user and multi-user applications, the minicabinet version works with many different displays and peripherals.

A rack-mountable version is available, too.



All are true 32-bit computers, with 32-bit CPUs, memories, and data paths. And the multi-CPU architecture lets you nearly double or triple your processing power at any time by adding one or two CPU boards. Without increasing the computer's size.

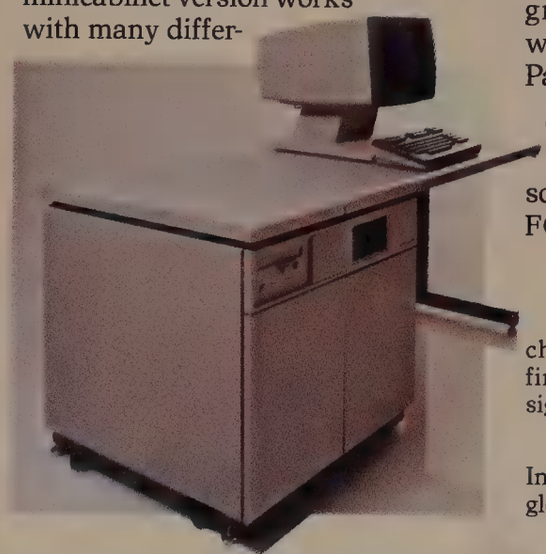
Two operating systems are better than one.

The integrated workstation is available with a choice of operating systems. One is HP's highly evolved, high-performance Enhanced BASIC, augmented with 3-D graphics and a software innovation called a run-time compiler. This substantially increases program execution speed, while retaining an interactive development environment.

The other operating system, called HP-UX, is a fully supported, extended version of the popular UNIX® HP-UX, available on all HP 9000s, adds virtual memory, graphics, data base management, data communications, and enhanced file capability to the basic UNIX 'shell.' High-level programming languages available with HP-UX are FORTRAN 77, Pascal and C.

Software, and plenty of it.

Much of the vast range of existing software written in HP BASIC, FORTRAN 77, Pascal and C is transportable to the HP 9000. HP will also be offering proprietary software packages emphasizing computer-aided design and engineering. These will tie the HP 9000 into HP's Manufacturer's Productivity Network (MPN). Third-party software suppliers will be providing many of the most widely used CAE packages for 32-bit computer systems. And both HP 9000 operating



As a minicabinet, it can handle multiple users.

The 32-bit CPU chip is bonded to the finstrate which doubles as a signal carrier and heat sink.

Up to three CPU boards and three Input/Output Processors can fit into a single HP 9000.

TCG-207

puter
s and scientists

systems offer extensive program development tools.

You also get a choice of communication tools. The HP 9000 is currently compatible with Ethernet,™ and with HP's Shared Resource Manager (SRM) which lets clusters of HP 9000 and 16-bit desktop computers share data and use common peripherals.

Links to central computers are also available. And in late 1983, HP will offer local area networks based on the IEEE-802 standard.

New technology from the silicon up.

The five superchips that make the HP 9000 possible are the 32-bit CPU, which can execute a million instructions per second; an eight-channel Input/Output processor (IOP); a random-access memory chip capable of storing 128K bits of data; a memory controller that 'heals' up to 32 bad memory locations; and an 18-megahertz clock.

Hewlett-Packard's advanced NMOS-III process makes it possible to put 450,000 transistors on a chip only 0.4 centimeters square. This tremendous density of electronic components could have required an expensive and elaborate cooling system.

Instead, HP engineers developed a new mounting structure called a finstrate, a copper-cored circuit board, which acts as both cooling fin and substrate. The finstrates containing the CPU, IOP, memory, and clock chips are housed in a lunchpail-sized module.

One user, one mainframe.

Clearly the trend in engineering and scientific computation is away from large machines shared by multiple users and towards networks of powerful personal workstations, sharing peripherals and data bases. The reason is compelling. An engineer or scientist in personal control of an HP 9000 can solve so many more problems more easily that the increased productivity alone makes the cost of individual computers easy to justify.

For complete information about this powerful breakthrough in 32-bit computing, contact the local HP sales office listed in your telephone directory. Ask a Technical Computer Specialist for a demonstration. Or write to Pete Hamilton, Dept. 56151, Hewlett-Packard, 3404 East Harmony Road, Fort Collins, CO 80525.

HEWLETT PACKARD

UNIX is a registered trademark of Bell Laboratories. Ethernet is a trademark of Xerox Corporation.

gene transfer mediated by transposons should make possible the modification of plant and animal genomes, for example to improve agricultural yield or to confer resistance to disease.

The Source at the Center

The center of the Milky Way galaxy is hidden by stars, gas and dust, but lately clues to what lies in the center have begun to emerge. One clue is that the galactic center emits gamma rays, the highest-energy form of electromagnetic radiation. Furthermore, the gamma-ray emission is marked by a sharp spectral line: all the rays have the same energy, namely .511 million electron volts. This is precisely the energy equivalent of the mass of an electron or a positron (the antiparticle of the electron), and so it is all but certain the gamma rays are generated in pairs when electrons and positrons collide and annihilate each other. The radiation from the galactic center was discovered in 1977, when a gamma-ray detector suspended from a balloon was flown over Alice Springs in Australia by a group led by Marvin Leventhal of Bell Laboratories and Crawford J. MacCallum of the Sandia National Laboratories.

Further observations now suggest that the gamma-ray emission varies with time; indeed, it seems to have stopped. A detector placed aboard the *HEAO-3* satellite by workers at the Jet Propulsion Laboratory shows that between September, 1979, and March, 1980, the emission declined by two-thirds, and the most recent measurement made by Leventhal and his colleagues (on November 21, 1981) shows no emission at all. The absence of emission is confirmed by investigators at the Goddard Space Flight Center of the National Aeronautics and Space Administration and the Saclay Nuclear Research Center in France, who launched a balloon-borne detector on November 20, 1981.

In *The Astrophysical Journal* Leventhal and his colleagues comment on the various observations. "It would appear," they write, "that the source was going into a 'low' state at the time of the 1980 March *HEAO-3* observations and has remained there through 1981 November." Moreover, "it seems safe to conclude that the time scale for change is at least as short as six months." The time scale serves to limit the size of the source. It is a "relatively compact object," considered on a galactic scale; apparently it can be no larger than one light-year, the diameter a shell of light would have if it expanded for six months from a central point. (The speed of light limits the speed at which any process spreading through space could cause a coherent change in emission.)

Is the source a black hole? Leventhal and his colleagues cite observations at infrared wavelengths showing the presence of clouds of ionized gas within 10 light-years of the center of the galaxy. The velocities of the clouds suggest they are orbiting a supermassive central object. "Apparently no problem exists" in hypothesizing that a variable flux of positrons boils off the disk of hot gas thought to ring a black hole. The positrons, passing through the surrounding clouds of ionized matter, would then have a good chance of colliding with electrons, yielding gamma rays.

Platonic Chemistry

Ever since Friedrich August Kekulé grasped the shape of the benzene molecule in a dream of snakes biting their tails, the topology and geometry of organic molecules has been one of the more charming pursuits of structural chemistry. The DNA molecule twists thousands of times before its ends can join to form a ring, and certain enzymes can tie DNA into topologically interesting knots. Only recently, however, has anyone been able to create the molecular analogue of a Möbius band, a strip joined end to end after a single half twist. Likewise, the orderly arrangement of atoms in a crystal can exhibit remarkable periodic symmetry, yet creating molecules with the form of the five Platonic solids has proved a considerable challenge. The synthesis of a molecular dodecahedron has now been reported.

The molecular Möbius band was synthesized by David M. Walba, R. Curtis Haltiwanger and Rodney M. Richards of the University of Colorado at Boulder; they made the band in much the way it is done with paper, scissors and paste. First carbon and oxygen atoms are bound together in two parallel chains, and the chains are fused by double bonds between carbon atoms lying opposite each other on the chains. The double bonds form the rungs of a ladder-shaped molecule called THYME diol ditosylate, and the network of bonds in the ladder defines a surface analogous to the surface of a long, narrow sheet of paper. The top and bottom of the ladder have a chemical affinity for each other, and so the ends can be joined, forming a ring. If the legs of the ladder cross when they join, the surface defined by the network of atomic bonds has the same topological properties as a strip of paper given a half twist and then pasted together at the ends.

One striking characteristic of the Möbius band is that it has only one continuous edge; run a finger along the edge of the paper model and the entire edge will be traversed before the starting point is regained. Moreover, slitting the band lengthwise around its entire circumference yields a single, thinner band that is twice as long and has two half twists. Walba and his colleagues carried out the chemical analogue of the slitting operation. When the molecule is immersed in ozone, the double-bonded carbon rungs of the molecular ladder are broken; the resulting molecule has indeed proved to be a ring twice the original size. The next step, according to Walba, will be to build longer molecular ladders so that more twists of the legs are possible before the ends are joined. Slitting a band with two half twists gives two separate bands linked together; slitting a band with three half twists gives a single band that winds through itself to form a trefoil knot.

The synthesis of the molecular dodecahedron is a task of even greater delicacy: the pieces of the 12-sided polyhedral cage must be attached in a strict sequence or the entire structure will collapse of its own energetic imbalance. Leo A. Paquette and his co-workers at Ohio State University are the only structural chemists to have achieved the synthesis so far, although other investigators are testing other routes to the same result. Paquette and his colleagues began with the hydrocarbon cyclopentadiene, made up of five carbon atoms arranged in a regular pentagon that eventually becomes one of the sides of the dodecahedron. In subsequent steps sides of the dodecahedron are added in such a way that the intermediate molecules remain roughly symmetrical about an axis. The final molecule, called dodecahedrane, is framed by 20 carbon atoms that make up the vertexes of the dodecahedron. Twenty hydrogen atoms, one attached to each carbon atom, lie outside the polyhedral surface defined by the carbon atoms and the carbon-carbon bonds.

The dodecahedron is the last of the five regular Platonic solids to be synthesized. Cubane, in which carbon atoms form the vertexes of a cube and a hydrogen atom is attached to each carbon atom, was made in 1964 by Philip E. Eaton and Thomas W. Cole, Jr., of the University of Chicago. The corresponding four-sided carbon skeleton would be called tetrahedrane; it has not been synthesized, but carbon atoms have been made to form a tetrahedron by investigators in Germany, who emplace butyl groups (C_4H_9) at the four vertexes. Both octahedrane, the eight-sided organic molecule, and icosahedrane, the 20-sided one, are probably too unstable to exist, but molecules in the two forms can be readily synthesized from atoms other than carbon.

Such geometric modeling may seem little more than an elegant trick, but the high degree of symmetry of the Platonic molecules makes it possible to investigate the properties of chemical bonds in great detail. Paquette also notes that a highly symmetrical molecule with an amino group attached to one vertex

readily passes through the membrane of a cell and tends to destroy virus particles inside it.

The Migrating Marsupial

When marsupial mammals began to flourish some 50 million years ago, they inhabited the New World (particularly South America), western Europe and Australia. The keen edge of Occam's razor makes it unlikely that the three groups evolved independently, but the question has remained whether New World marsupials had colonized Australia, as they evidently had Europe, or whether the order originated in Australia and radiated from there. The further question of how the animals could have migrated in either direction was essentially answered some years ago by students of continental drift, who found that when the marsupials were developing, South America was connected to Australia by way of Antarctica.

Until recently, however, the Antarctic fossil record, which is surprisingly rich in fishes, reptiles, birds and even sea mammals, gave no evidence of marsupials having passed that way either coming from or going to Australia. Such is no longer the case. Writing in *Science,* Michael O. Woodburne of the University of California at Riverside and William J. Zinsmeister of Ohio State University have announced the discovery of fossil marsupial jawbones and teeth of late Mesozoic to early Cenozoic age in a sandstone formation on Seymour Island, off the Antarctic Peninsula.

The fossils, the first evidence of land mammals found so far in the Antarctic, have been assigned to the extinct marsupial family of polydolopids. Remains of the group were unearthed previously in Patagonia, Bolivia and Brazil. Because polydolopids had not been known from anywhere except South America, their presence in the Antarctic is strong evidence that they came there from the New World rather than from Australia. On the other hand, late Mesozoic and early Cenozoic marsupial fossils from Australia are unknown, and later ones all resemble one or another of the eight marsupial families that still survive there rather than any potentially ancestral foreign marsupial. Hence the issue of which way the Mesozoic migration went is still unresolved.

Unforgettable Data

In the earliest electronic computers space for the storage of information was a precious resource; as a result programs were written in the most compact form possible. As the cost of computer memory has decreased, the need for succinctness in programs has diminished. More emphasis can now be put on other qualities, such as the speed at which a program operates or the ease with which it can be written, understood and altered. The effects of further decreases in the cost of memory have recently been discussed by George P. Copeland of the Servio Logic Corporation in Portland, Ore. Writing in *Computer,* he describes the likely direction of events by speculating on what a computer system would be like if the mass storage of information were completely free.

Mass storage, which commonly takes the form of magnetic tapes or disks, is the cheapest kind of computer memory. Large quantities of information, such as the payroll and employment records of a business, can be kept indefinitely in a mass-storage device; when a particular record is needed, it is called up into the internal memory of the computer. In the way such a "data base" is maintained today, records are periodically added, altered and deleted. The most important change that would be brought about by free mass storage, Copeland finds, is that no information would ever be deleted from the system.

The ability to delete information might seem to have certain advantages even apart from the clearing of space for new data. In particular, when an error is discovered, the usual means of correcting it is to delete the erroneous information and replace it with a revised version. Copeland points out that there are hazards inherent in this practice. For example, an error in stating an employee's salary could be corrected by the deletion-and-replacement method, but in the meantime the incorrect amount might have been used in calculating pension benefits, insurance premiums, tax withholding and so on.

Forbidding deletion eliminates such attempts to "go back into the past and change history," but a new awkwardness is introduced. Each time a record is modified a new version is created, but the old one also remains in the system; hence multiple copies of the record accumulate. Copeland proposes to keep track of the proliferating data by assigning each record two identifying numbers. One number is called a surrogate, and it remains constant for all versions of the record; the other number serves as a "time stamp" indicating when the version was created. Copeland notes that the nondeletion policy, in which errors can be adjusted but never expunged, is not a new idea in record keeping. It is an established practice in accounting, where "no journal entry or ledger posting deletes any information."

The cost of mass storage in computer systems will never decline to zero, but it may well become low enough to make the permanent retention of information feasible. Indeed, there is at least one existing low-cost storage medium that seems to require a nondeletion policy: the optical disk. On such a disk information is recorded as a pattern of microscopic pits etched into the surface by a laser beam. The inability to erase the pits has occasionally been cited as a disadvantage of optical-disk recording, but in Copeland's view it may be a virtue.

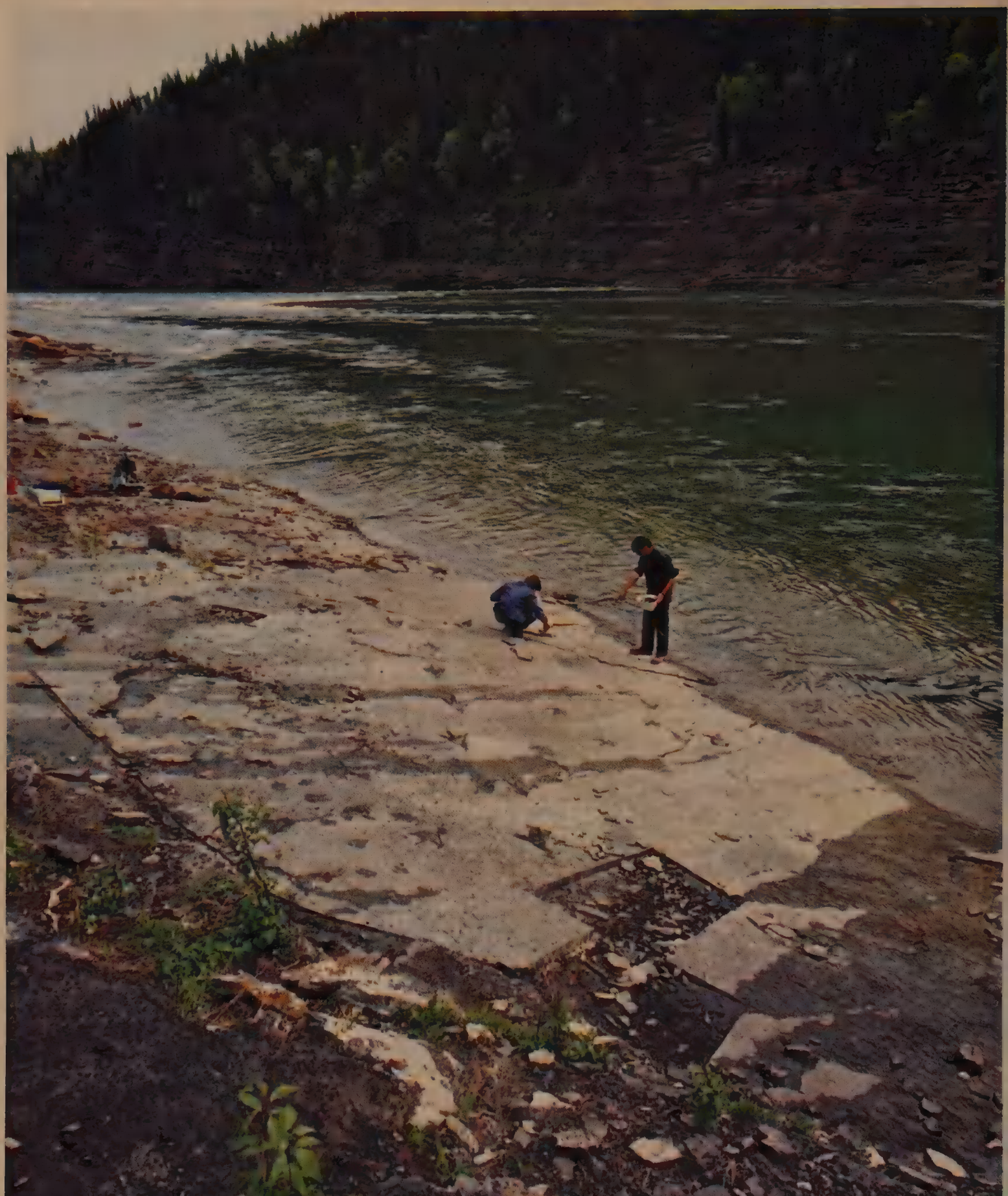
Natural Gas

The smokestacks of the world are viewed with concern in some quarters because the carbon dioxide and other gases they release into the atmosphere might inhibit the escape of heat into space. The concern is put into a new perspective by the recent report that the smokestacks are outgassed by termites, which munch huge amounts of biomass and give off various gases as by-products. Termites may contribute twice as much carbon dioxide to the atmosphere as smokestacks do, and the insects also emit significant amounts of methane (CH_4), molecular hydrogen (H_2) and other gases that are trace components of the atmosphere. The termite emissions are discussed in *Science* by P. R. Zimmerman and J. P. Greenberg of the National Center for Atmospheric Research, S. O. Wandiga of the University of Nairobi and P. J. Crutzen of the Max Planck Institute for Atmospheric Chemistry in Mainz.

The authors estimate the world's termite population at 2.4×10^{17} and calculate that the insect's efficiency of digestion is 60 percent or more. Digestion depends primarily on the decomposition of food by microorganisms in the termite's gut. On the basis of measurements made on termite colonies maintained in the laboratory the authors estimate that the gross amount of carbon dioxide produced by termites worldwide is 4.6×10^{16} grams per year, "more than twice the net global input from fossil-fuel combustion." The corresponding amount for methane is 1.5×10^{14} grams and for molecular hydrogen 2×10^{14} grams.

Zimmerman, Greenberg, Wandiga and Crutzen point out that the ecological impact of termites, as indicated by the ratio of what they eat to the net primary productivity of biomass, is greatest for wet savanna, Temperate grasslands, cultivated land in underdeveloped countries and areas that have been cleared and burned.

The authors caution that their estimates are not definitive. The laboratory measurements have an uncertainty of 50 percent and additional uncertainty arises from the possibility that not all termite species yield gases at the rates measured in the laboratory. Moreover, in a natural setting microorganisms in the soil may consume or modify some of the emissions by termites. "We estimate that all of these uncertainties may cause the actual gas productions to vary from those reported here by a factor of two."



FOSSIL-FOOTPRINT BONANZA, uncovered during five years of salvage work in advance of flooding at the Peace River Canyon dam site in British Columbia, included this array of dinosaur trackways preserved in early Cretaceous sediments. In the foreground one rectangular block of footprints has been removed. In the background investigators from the Natural History Museum of Alberta [redacted] preparing casts of other prints. Over four miles salvage workers have found 1,700 footprints representing at least 10 different dinosaur species.

The Footprints of Extinct Animals

Vertebrate animals have left their tracks in sediments ever since they first appeared on dry land 370 million years ago. Indeed, most of the known extinct species are known only from their footprints

by David J. Mossman and William A. S. Sarjeant

The skeletons of extinct land animals in museums suggest that the main fossil evidence for such animals is bones. This is not the case. The bones of the animals are very much rarer than their tracks. Indeed, many extinct land animals are known only from their tracks; their bodily remains appear never to have been preserved. In England, for example, the middle Triassic sediments of Worcestershire and the late Triassic sediments of Nottinghamshire, both laid down in the Mesozoic "age of reptiles," contain no fossils of vertebrate animals. Footprints in these same formations, however, indicate that at least eight different kinds of reptiles traversed the sediments of Worcestershire and six other kinds did the same in Nottinghamshire.

Tracks similarly testify to the earliest known invasion of dry land by vertebrates. The markings are found in the Old Red Sandstone formation of the Orkney Islands, of middle Devonian age (370 to 360 million years ago). They record the passage of a primitive fish. Like today's lungfishes, it lived at the mercy of seasonal dry spells. It left its track when it crawled on its belly overland from one pool of water to another, propelled by its stubby lobate fins.

From the perspective of the present the Devonian period may seem remote. Within the absolute-time framework of earth history, however, it is not so long ago. The point is readily appreciated if one views the 4.5 billion years that have passed since the earth formed as being a single year, with each day lasting for 12.3 million years. On such a time scale the earth's first forms of life—primitive plants resembling modern single-celled algae—appeared in the seas in early May. Many-celled forms of life, however, did not arise until early November.

By about November 20 primitive fishes were swimming in the planet's waters. Toward the end of the month their descendants ventured onto the land. By December 7 reptiles had become the dominant terrestrial animals, and by mid-December the first mammals had appeared. At about 5:00 P.M. on the last day of the year two early hominids left their footprints in a fresh fall of volcanic ash on the Laetoli Plain of Kenya. Our own genus, *Homo,* did not appear until about an hour before midnight—some 500,000 years ago. Thus the entire span of vertebrate life on land occupies less than six full weeks of an earth-history "year." Even though that important interval spans some 360 million years, it is a very small part of the history of the planet.

The Terrestrial Record

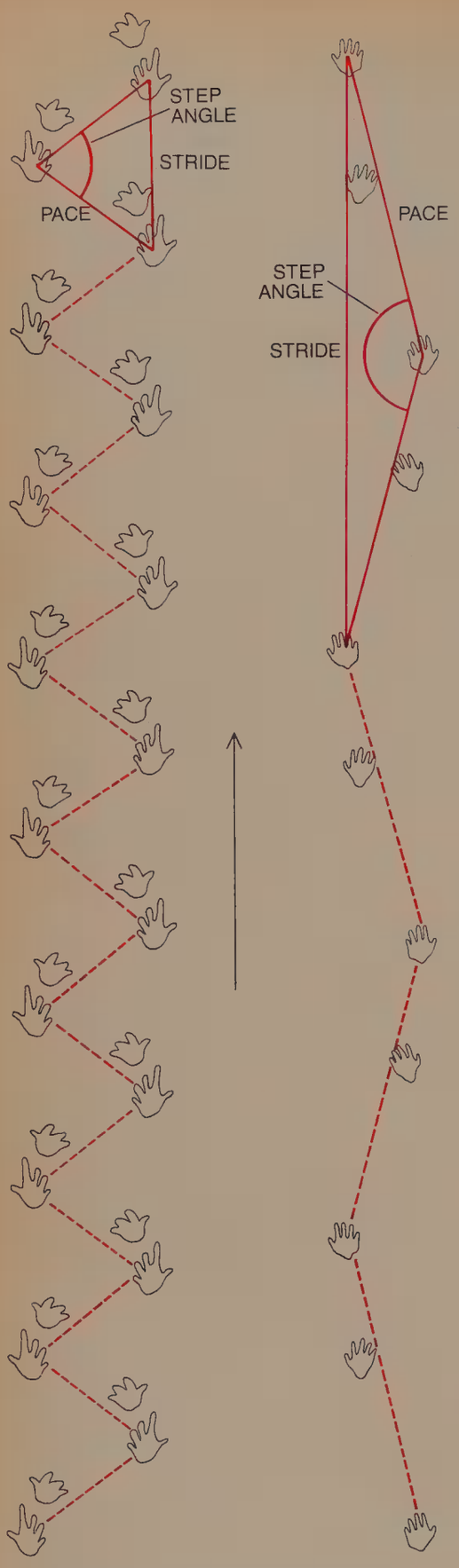
Throughout this interval two kinds of evidence were preserved that help paleontologists to interpret the animal life of the past. The remains of the animals themselves are evidence of one kind: fossilized bones, articulated or separated, with or without associated structures such as teeth, scales or feathers and, in rare instances, with fossilized impressions of skin or other tissues. From these remains one can reconstruct the animal's appearance with fair accuracy and deduce how it moved about and fed. Yet an animal has only one body and its remains, if they are preserved at all, can be found in only one place. The other kind of evidence is the animal's tracks and trails, which are a much more extended and dynamic testament. Studying these clues to an animal's behavior, the paleontologist is able, as it were, to see the animal in action.

The lobe-finned fishes that first ventured onto the land in Devonian times gave rise to four-legged amphibians. In evolutionary terms the transition was quick: the earliest amphibian tracks are found in formations of the late Devonian. Amphibian trackways are characteristically broad and show the short stride that is typical of rather inefficient quadruped locomotion. The individual prints show a full impression of the palms (indicative of a slow walking pace) and clawless digits that spread widely and are directed outward rather than forward. Amphibian footprints are abundant in the sediments of the Carboniferous period, which followed the Devonian. They are less abundant in the Permian (280 to 225 million years ago) and are rare in the fossil record thereafter, largely because the surviving amphibians preferred, as they still do, wet conditions that are unsuitable to the preservation of tracks.

The rise of the reptiles to dominance is clearly seen in the later Carboniferous and in the Permian as the number of their footprints steadily increases. Both the diversification of these animals and their progressively efficient locomotion are made evident by changes in the form of the digits, including the development of a short, thumblike fifth digit and the appearance of claws. In some reptilian lineages the trackway narrows and the stride lengthens. The animals were traveling faster but their gait remained quadrupedal. Other tracks show that by late Permian times some small reptiles maintained a quadrupedal gait when they moved slowly but adopted a bipedal one when they moved quickly. Their body pivoted at the hips, a long tail counterbalancing their forelegs and nearly erect trunk.

This was the reptile lineage that gave rise to the dinosaurs at the start of the Mesozoic era, 225 million years ago. Many Triassic dinosaurs were largely or entirely bipedal. Even in the groups that later resumed quadrupedal locomotion there was a marked disparity in limb size. The body continued to pivot on the pelvis and the hind feet were usually much larger than the forefeet. Of all fossil footprints dinosaur tracks are the most spectacular; they are found in abundance in terrestrial sediments of Mesozoic age in most parts of the world.

In time two groups of small dinosaurs evolved wings. In one group, the pterosaurs, membranes that enabled the animals to glide stretched outward from the flank of the body, supported by an elongated fourth digit of the forelimb. W. L. Stokes of the University of Utah has found fossil trails of these flying reptiles in sediments of the Morrison for-



TRACKWAY MEASUREMENTS appear in this diagram. At the left is the broad trackway characteristic of an amphibian; it shows the successive front and rear footprints of one of the amphibians that became extinct near the end of the Paleozoic era. As the nearly equilateral colored triangle indicates, both the pace and the stride of the animal were short. At the right is the narrower trackway characteristic of a reptile; it shows the successive front and rear footprints of a dinosaur. Both the pace and stride distances are proportionately much longer than those of the amphibian, an indication of more efficient locomotion.

mation, a late Jurassic deposit in Arizona. The trackways make it clear that the animals not only were capable of gliding but also could walk readily on all fours, with the wing membrane trailing on the ground behind the forefeet.

In the second group a modification of the epidermal scales led to the evolution of feathers and a more efficient means of flight. The evolution of birds from dinosaurs produced no immediate modification of the hind feet, but footprints distinctly avian in style were not long in appearing. The transition is evident in the beautifully preserved footprints of birds found in the middle Cretaceous sediments of eastern British Columbia by Philip J. Currie of the Tyrrell Museum of Paleontology in Drumheller, Alberta.

Not all the Mesozoic reptiles became bipedal. The synapsids were one persistently quadrupedal group, a reptilian lineage that was separated from the dinosaurs throughout their history. Here too, however, the stride of the animals lengthened, the trackway narrowed and the digits were directed forward, all changes that are evidence of increasingly efficient and fast locomotion. The structure of the synapsid foot became increasingly like the structure of the foot of a mammal. Indeed, in the sediments of late Triassic and early Jurassic age in the province of São Paulo the Brazilian paleontologist Giuseppe Leonardi has found numerous tracks so exactly intermediate in form between reptile and mammal footprints that they could be attributed to either group.

With the end of the Mesozoic dinosaur tracks vanish from geologic history. From the beginning of the Cenozoic era, 65 million years ago, the variety of mammal tracks corresponds to the increasing diversification of these advanced vertebrates. That diversification is one of the main evolutionary events of the Tertiary period, the first and longest part of the Cenozoic. It seems only fitting that with the succeeding Quaternary period, as geologic history comes up to the present, we find among the mammal tracks at Laetoli evidence that some of man's upright precursors have walked their way into the fossil record.

The Trail of Discovery

What is the earliest documented discovery of fossil vertebrate footprints? One might think it was made in Europe by one of the pioneers of geology, but it was not. The locale was the valley of the Connecticut River near South Hadley, Mass., where red sandstone beds of the late Triassic to the earliest Jurassic age lie exposed. The discoverer was Pliny Moody, a local farm boy, the year was 1802 and the tracks were those of small dinosaurs. At that time, however, dinosaurs were unknown and so the footprints were taken to be those of large birds. Some even went so far as to suggest that they were the tracks of Noah's raven.

A quarter century passed before a scholarly account of fossil footprints appeared. The footprints were in a sedimentary formation of Scotland, the New Red Sandstone. The age of the formation was then uncertain, but it is now known that the sediments were deposited in the final interval of the Paleozoic era. The discoverer was a local clergyman and amateur naturalist, the Reverend Henry Duncan. His description of the tracks, published in 1828, astonished the scientific world of the time.

The prints, which show the movement of an animal with a short stride and a broad trackway, are now recognized as those of a caseasaur, one of a group of large, ponderous plant-eating reptiles. Like the dinosaurs of the Connecticut River valley, the caseasaurs were still unknown, and so the tracks were attributed, quite reasonably, to some kind of turtle. Having learned of Duncan's discovery, a distinguished English geologist, William Buckland, undertook to test the turtle-track hypothesis. John Murray III, a well-known publisher, described Buckland's investigation:

"I went on Saturday last to a party at Mr. Murchison's house, assembled to behold tortoises in the act of walking upon dough. Prof. Buckland acted as master of the ceremonies. There were present many other geologists and savants, among them Dr. Wollaston. At first the beasts took it into their heads to be refractory and to stand still. Hereupon the ingenuity of the professor was called forth in order to make them move. This he endeavored to do by applying sundry flips with his fingers upon their tails; deil [devil] a bit however would they stir; and no wonder, for on endeavouring to take them up it was found that they had stuck so fast to the piecrust as only to be removed with half a pound of dough sticking to each foot. [It] was found necessary to employ a rolling pin, and to knead the paste afresh; nor did geological fingers disdain the culinary offices. It was really a glorious scene to behold all the philosophers, flour-besmeared, working away with tucked-up sleeves. Their exertions, I am happy to say, were at length crowned with success; a proper consistency of paste was attained, and the animals walked over the course in a very satisfactory manner; insomuch that many who came to scoff returned rather better disposed toward believing."

After Duncan's discovery numerous other footprints began to be reported but not necessarily accepted. For example, in 1841 on the shores below Horton Bluff in Nova Scotia, William Logan, a visiting Scotsman, found a slab of sedimentary rock bearing well-pre-

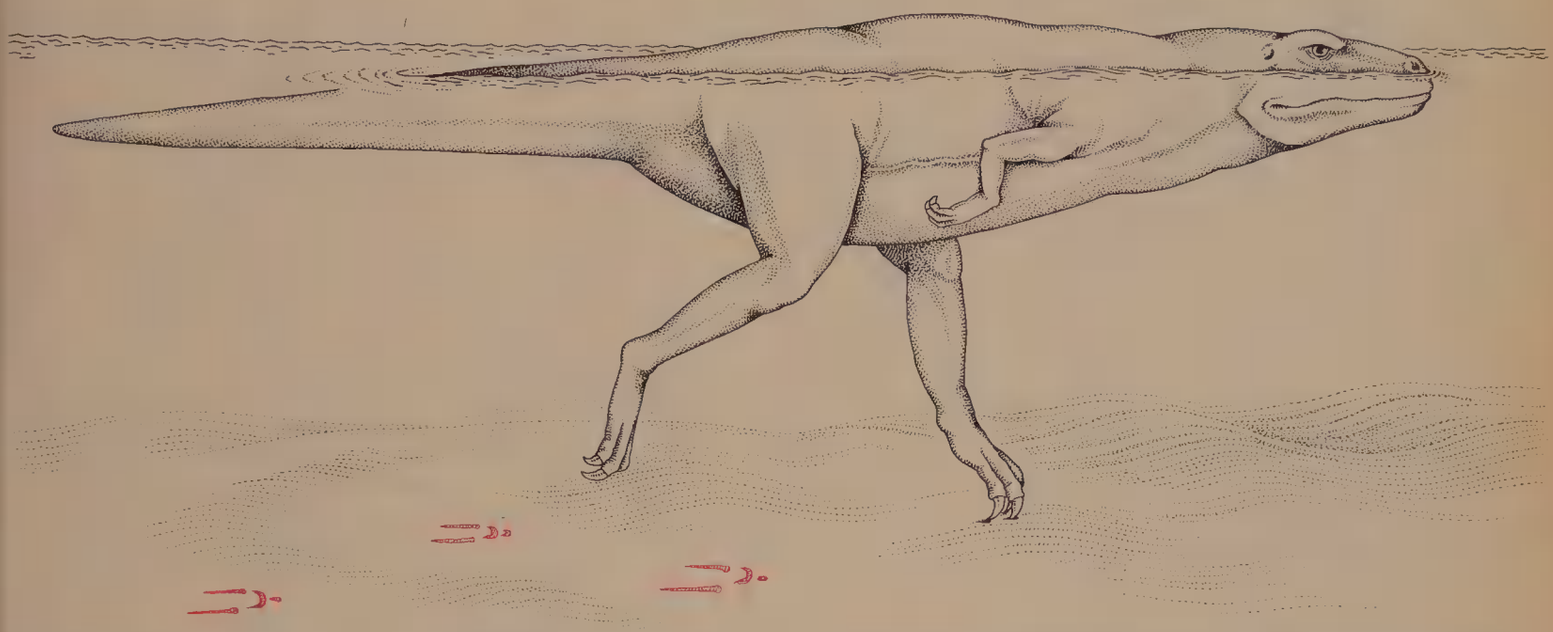
served vertebrate footprints. The formation was Carboniferous in age, a time, in the prevailing opinion of Logan's day, too early in geologic history for any vertebrates other than fishes to have existed. When, on returning to Britain, Logan exhibited his find to the Fellows of the Geological Society, the evidence was disbelieved.

Logan was not vindicated until 1872. In that year John W. Dawson, working with a founding father of geology, Sir Charles Lyell, was investigating Carboniferous strata on the east coast of the Bay of Fundy. The strata included the coalfields at Joggins in Nova Scotia, where a number of upright tree stumps had been preserved as fossils. Dawson had predicted that the fossil remains of small amphibians and reptiles might be discovered in the hollow interior of the stumps. His prediction was correct; reptile remains were indeed found inside the stumps. Perhaps the animals had been trapped in the stump hollows while searching for water or insects. In any event it was now clear that the Carboniferous fauna included terrestrial vertebrates as well as fishes.

Logan and Dawson were destined to be major figures in Canadian science. Logan was the first director of the Geological Survey of Canada and Dawson became principal of McGill University. Both men were eventually knighted.

A Waning Interest

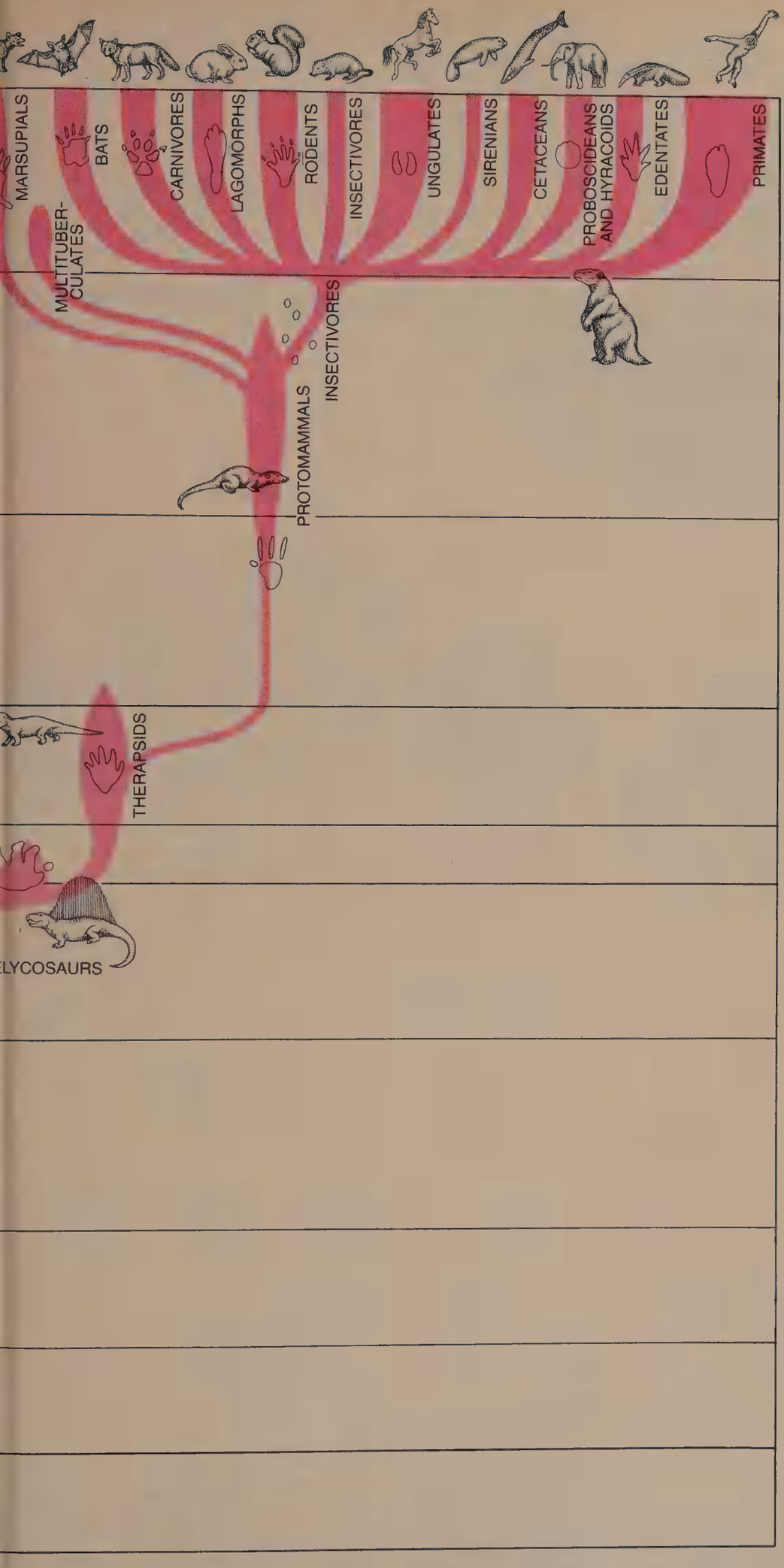
Although many descriptions of fossil vertebrate tracks were published in the second half of the 19th century, such papers appeared in diminishing numbers in the first half of the 20th. This branch of paleontology, known formally as ichnology (from the Greek *ikhnos,* track), fell progressively into disfavor among geologists. The reasons are not hard to find. Between 1930 and 1960 the



SWIMMING TRACKS of a carnivorous dinosaur found recently in sediments of Jurassic age consist of clear toe and claw marks that give no indication of foot pressure on the bottom. The depiction of the swimming carnivore is based on a restoration of *Megalosaurus* by Matthew Hyman. The tracks indicate that herbivorous dinosaurs, the carnivores' natural prey, could not escape by going into the water.

DINOSAUR TRACKS in Connecticut River valley red sandstones of early Jurassic age indicate that the animal was essentially bipedal, as at the left, but that sometimes it walked on all fours, as at the right. This is one of the many extinct reptiles known only from their tracks; from these Richard Lull of Yale University reconstructed the animal as it is seen here. It was one of the early ornithopods, a herbivore.

VERTEBRATE EVOLUTION has contributed numerous tracks to the fossil record, including the earliest terrestrial trackway: that of crossopterygian fish left in the Old Red Sandstone of the Orkney Islands in middle Devonian times (*fifth band from left*). Thereafter,



the representative footprints indicate, a succession of amphibians, reptiles, birds and mammals have left their traces. The last print, at the far right, was made by a hominid in Kenya.

few papers in the field were written primarily by amateurs rather than professionals, appeared in relatively obscure journals and were poor in quality. The authors' descriptions of the tracks they had discovered were woefully inadequate and their findings were seldom related in any systematic way to tracks that had been recorded previously. Photographs of tracks, if they were included at all, generally were made with overhead illumination rather than with the side lighting needed to bring out shallow impressions or the low relief of casts. More often the illustrations were not even poor photographs but were inadequate line drawings. The authors almost never stated where the specimen could be examined by others, chiefly because most of the specimens were never placed in study collections. Here the fault was not necessarily the authors'. Since footprint slabs are large and awkward to handle, they tend to be unpopular with museum curators. After the slabs are received they are likely to end up outside the museum, either flat on the ground, where they are worn away by the feet of visitors, or propped up, where exposure to the elements eventually destroys them.

In the 1950's the investigations of the German paleontologist Walter H. Häntzschel and the missionary endeavors of his disciple Adolf Seilacher resulted in a surge of interest in the fossil tracks left by marine organisms. Even then there was no parallel revival of interest in the tracks of vertebrates. Today only a handful of investigators are concerned with them, most notably Leonardi and Rodolfo Casaniquela in South America, Donald Baird of Princeton University, Justin Delair in Britain, Georges Demathieu and Albert F. de Lapparent in France, Hartmut Haubold in East Germany and O. S. Vialov in the U.S.S.R.

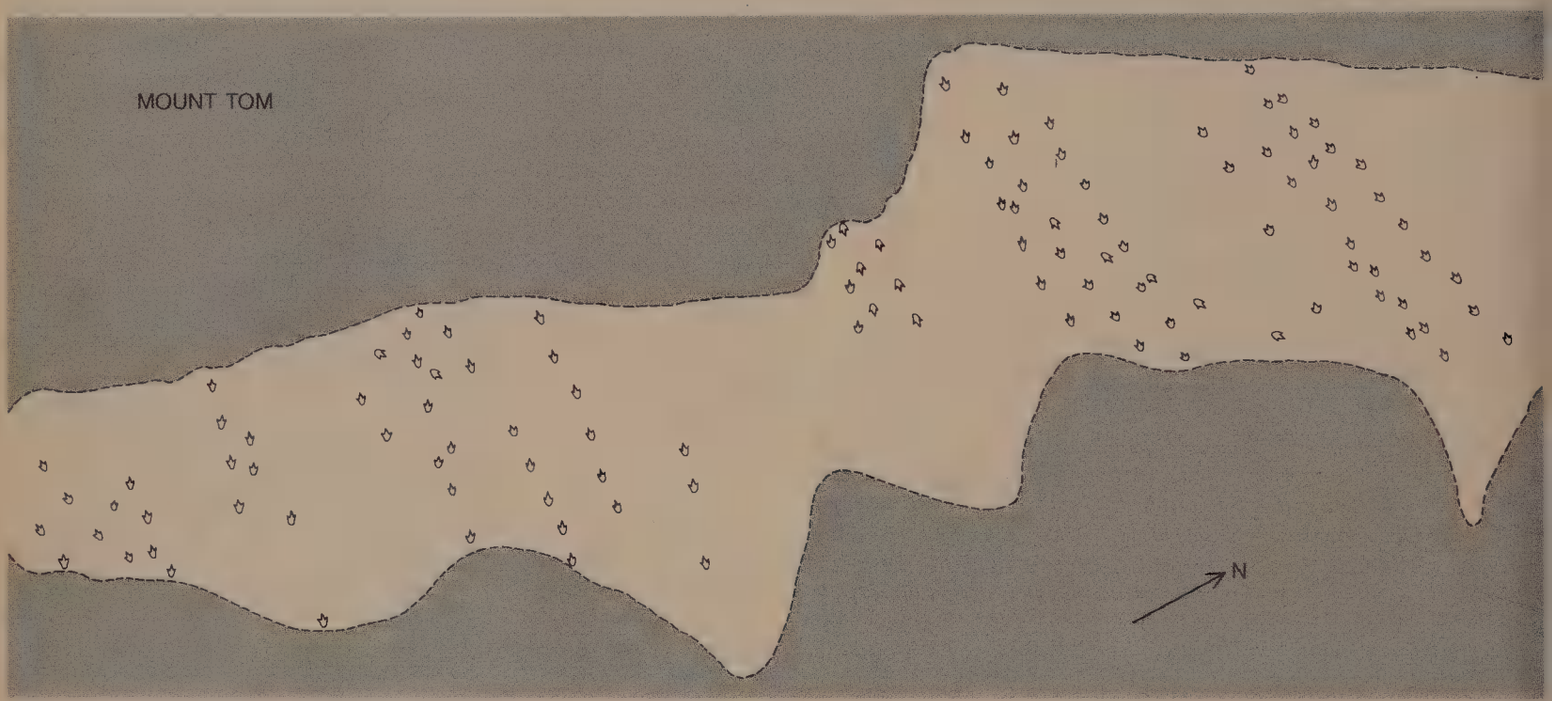
In addition to the problems of an inadequate scholarly literature the student of fossil footprints is confronted by other difficulties. Foremost among them is the initial character of the tracks themselves. Ideally for clear foot impressions to be formed there first must be a moist, fine-grained and cohesive bed of sediment for the animal to traverse. This it should do slowly, leaving detailed impressions of its forefeet and hind feet (assuming it is quadrupedal). Under such ideal circumstances even the exact outlines of claws or nails, the shape of pads and the pattern of scales may be preserved.

Actually few such high-quality impressions are found. If the sediment is too coarse, it will not retain details. If it is too wet, its deeper hollows will fill with water, distorting the shape of the print. If it is too dry, it will not be cohesive enough to preserve the impression. A strong wind may obliterate the print,

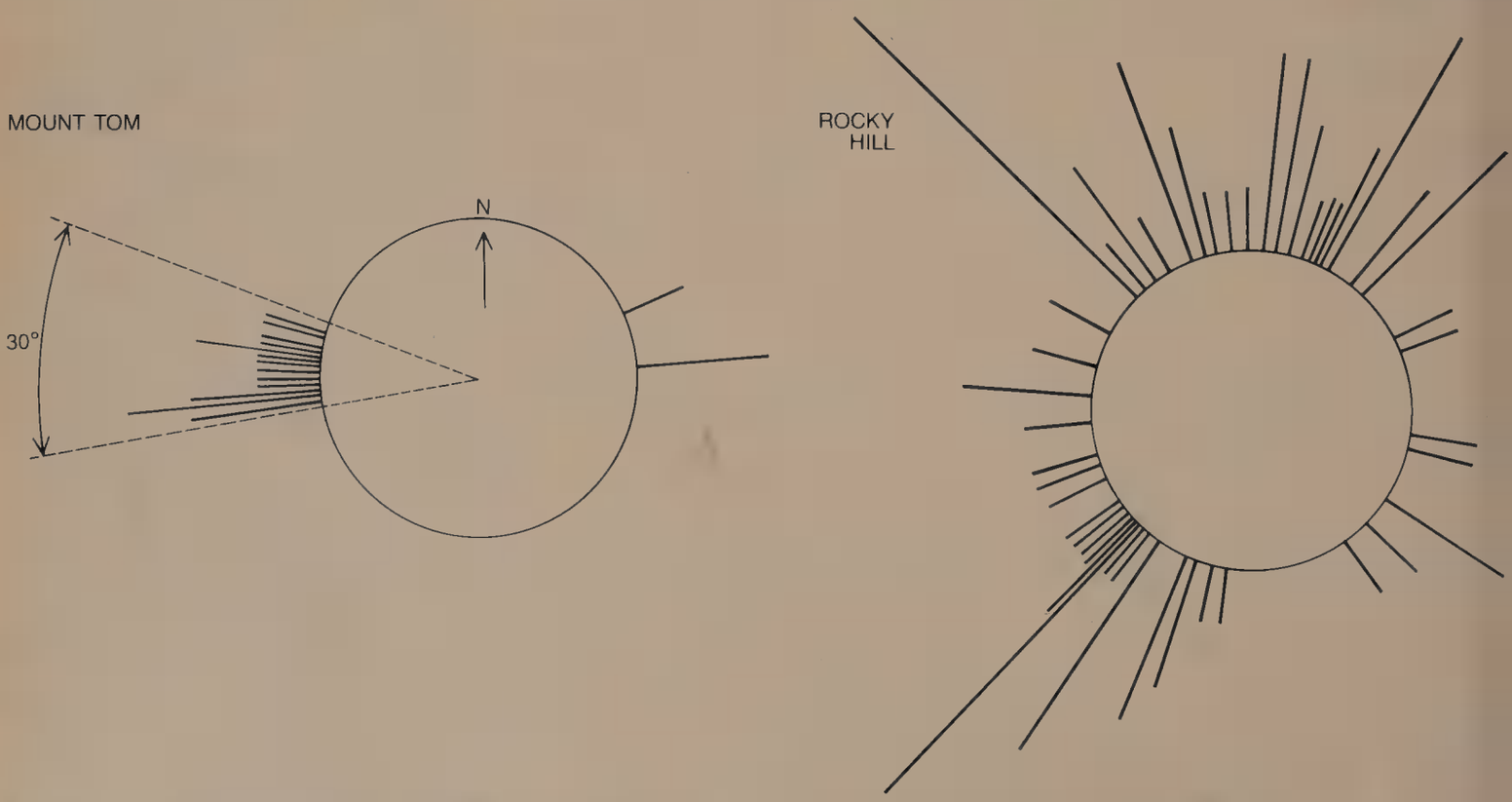
and if it was made near the seashore, a rising tide may do the same. Even the deposition of sediment on top of the tracks, which is clearly crucial to their preservation, may mean their obliteration. If the new sediment is too much like the imprinted one, the two strata will tend not to separate and so the tracks may never be discovered.

The most favorable conditions for track preservation exist after high waters have receded, leaving fine-grained sediment freshly exposed. At the seashore this means after the highest level of a spring tide. Inland it means after a rainy season has raised the stream or pool to its highest level. Imprints in such a moist surface will solidify as the sediment dries and will be buried by the next deposition of sediment. Even then it may be not the imprint that is preserved but rather its cast, particularly if the newer sediment is coarser and more cohesive than the older one. Such casts must be sought on the underside of sedimentary layers and may escape detection unless the layer has been turned over by the collapse of a cliff, say, or in the course of a commercial quarrying operation.

Lastly, the probability that a terrestrial animal track will be preserved and later recognized is directly proportional



SOCIAL BEHAVIOR among dinosaurs is suggested by the orientation of these fossil trackways of Triassic age in the Connecticut River valley near Holyoke, Mass. Analysis of the tracks by John H. Ostrom of Yale reveals that at least 19 carnivorous dinosaurs of the genus *Eubrontes* had crossed damp sand traveling from east to west. The common direction suggests that the animals were hunting in a pack.



ROSE DIAGRAMS contrast the nonrandom orientation of the *Eubrontes* tracks, at the left, with the far more random orientation of more than 1,000 similar dinosaur footprints at another Connecticut River valley site, Rocky Hill, at the right. Not all the recorded tracks at either site are shown. At the left, however, 70 percent of all the trackways lie within a 30-degree arc. The randomness evident at the right suggests that the Rocky Hill trackways were imprinted over a longer period of time, but this does not rule out gregarious activity.

to the animal's size. Large animals need to range widely for food and to visit water holes frequently. Moreover, their footprints are more deeply impressed and therefore less readily destroyed. Although small animals are much more numerous than large ones, they range less widely, require less food and water and leave shallower footprints. That is why a dinosaur's tracks are more likely to be preserved (and far more likely to be recognized) than a lizard's.

The Analysis of Tracks

Ideally a trackway suitable for analysis should consist of a sequence of at least three footprints or casts. Among quadrupeds locomotion is initiated with the hind foot on one side, followed by the forefoot on the same side and then by the hind foot and the forefoot on the other side. In fast locomotion two or three feet touch the ground simultaneously; in slower locomotion three feet or all four do so. In jumping locomotion all four footprints are set close together; they do not overlap. Jumping tracks are extremely rare in the fossil record; some were recently found by Leonardi and one of us (Sarjeant) in a Brazilian formation of Mesozoic age. The tracks of bipeds normally show the left and right feet alternating; the footprints are rarely side by side.

The analyst's first job is to make four basic measurements of a trackway. The first is a measurement of stride, or forward movement; it is made from a fixed point on one footprint to the same point on the next print of the same foot. The second is a measurement of pace; it is the distance between the right forefoot and the left forefoot and between the right hind foot and the left hind foot. The length of the stride is usually identical for the forefeet and the hind feet, but the length of the pace may differ greatly. The third is a measurement of pace angulation, or step angle; it is the angle formed by joining the midpoint of three successive hind-foot prints or three successive forefoot prints. The fourth is a measurement of trackway breadth. As we have noted, the mark of an inefficient walker is a broad trackway and a short stride. Conversely, a narrow trackway and a long stride are evidence of an efficient walker moving fast. A trackway moderate in both width and stride, however, can be evidence of either a not too efficient walker or an efficient walker moving at a relatively low speed.

For quadruped tracks a further important measurement is one from the midpoint between two consecutive hind-foot impressions to the midpoint between two similar forefoot impressions. This distance is actually between the hipbone socket and the shoulder socket of the living animal. It corresponds approximately to the length of the animal's trunk (head, neck and tail omitted).

Because the margins of both the footprints and the casts may grade smoothly into the surrounding matrix exact boundaries are often difficult to determine. As a result inaccurate measurements of footprints are all too common and the field investigator must guard against them. A further hazard arises from the fact that in some instances the weight of the animal making the track was impressed not only on the surface layer of the sediment but also on layers farther down. Care must therefore be taken to distinguish between an actual footprint and a "ghost print": a subsurface impression. Ideally the investigator should examine not only the track but also a cast of it. If what has been found is itself a cast, he should examine a mold made from it (which, of course, re-creates the missing track). If it is a true footprint (a mold), then a cast should be made of it. Such casts are made by applying plaster of Paris, latex or fiberglass to the footprints in the field. They provide a valuable record for subsequent study.

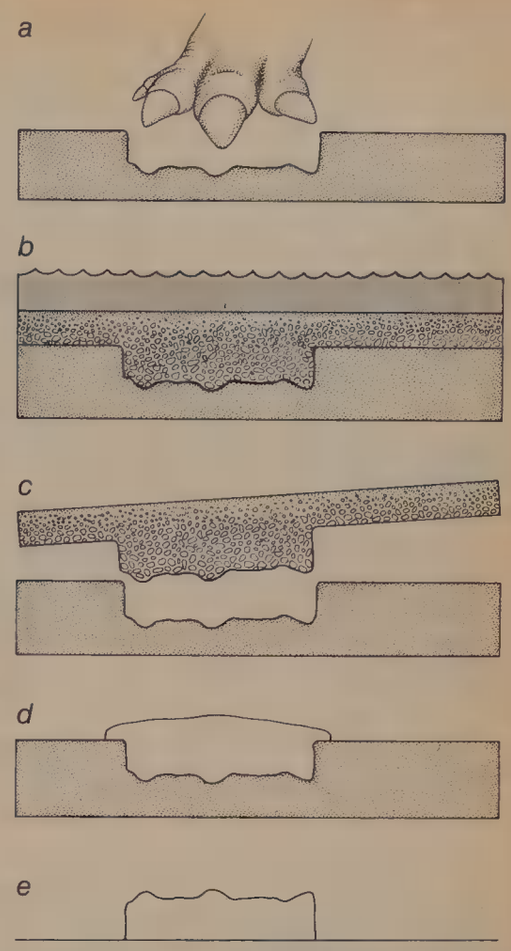
Another aid to track analysis is the measurement of the degree of overlap between the forefoot and hind-foot impressions. With the exception of certain flexible-bodied animals such as salamanders, the extent and consistency of overlap yields a second measurement of the distance between the quadruped's hip and shoulder. Still another useful measurement determines the position of the footprints with respect to the midline of the trackway; it can yield information on a quadruped's gait. For example, if the gait is a sprawling one, the midline of the trackway will not correspond exactly to the midpoints between paired footprints.

Reconstructing Unknown Animals

The hazards of attempting the reconstruction of an animal known only from its fossil footprints are well demonstrated by a 19th-century example. The animal is *Chirotherium,* from the Greek roots for "hand animal." Its tracks were first discovered in 1834, imprinting a red sandstone in the central German state of Thuringia dating to the Triassic (from 225 to 195 million years ago). The hind-foot prints are about the size of a man's hand; four of the digits are directed forward but the fifth is set at an angle to the other four, much as the thumb of the human hand is. Some hind-foot prints were associated with smaller but very similar impressions that were taken to be forefoot prints.

Even though the dominant Triassic land animals were reptiles, the Thuringian fossil footprints were at first attributed to some kind of ape or bear. What blinded those who arrived at this conclusion was a misinterpretation: they identified the angled digit as a thumb even though it was on the outside of the print. Indeed, some early reconstructions pictured *Chirotherium* as a toadlike amphibian that walked cross-legged; at least this served to put the "thumb" in a normal position. Addressing the British Association for the Advancement of Science in 1874, the zoologist L. C. Miall demonstrated that the cross-legged reconstructions were false and proposed instead that the tracks were those of dinosaurs. That the tracks were at least reptilian became evident when it was recognized that the *Chirotherium* trackways were narrow. Amphibian trackways are invariably broad.

The most convincing evidence for the tracks having been made by a reptile was the fact that the best-preserved *Chirotherium* prints showed the impressions of scales. Moreover, they had strong claws, suggesting the reptile was carnivorous. Track analysis suggested the body of the animal was about a meter long and, since the gait was predominantly bipedal, was probably balanced in an upright position by a tail of equal length. Making allowance for the ani-



FOOTPRINT PRESERVATION requires, once a print has been made (*a*), that the indented sediment be filled with sediment of contrasting texture (*b*) and that both sediments harden. The two sediments may later split apart. If only the filling sediment is preserved (*c*), it is a natural cast. If only the indented sediment is preserved (*d*), it is a natural print and can be used to make casts (*e*).

BIRD FOOTPRINTS appear scattered across an exposure of Mesozoic sediments in British Columbia. This is the famous Peace River Canyon site, where large numbers of dinosaur tracks were discovered in 1922. The shore birds that left these tracks are otherwise unknown in the fossil record. Their footprints are the only traces of bird life found in early Cretaceous strata.

GROUND-SLOTH TRACK of elephant-size footprints crosses this exposure of the Río Negro formation in Argentina. The track was made in late Pliocene times, 3.5 million years ago.

mal's head and neck, this gave *Chirotherium* a length from the tip of its snout to the tip of its tail of about two and a half meters.

It is now known the animal that made the tracks was one of the pseudosuchian reptiles, an important early group ancestral to both the dinosaurs and the crocodiles. Although nearly two dozen species of *Chirotherium* are now recognized on the basis of footprint form, only a few fossil skeletons of corresponding types are known. One of them, *Ticinosuchus,* discovered less than 20 years ago in a formation of middle Triassic age in Switzerland, is a prime candidate for being the animal responsible for the tracks that puzzled the 19th-century paleontologists.

Just as the pattern of scales in certain *Chirotherium* footprints helped to clinch the animal's identity as a reptile, so the prints of other animals can reveal details of soft parts that are only rarely fossilized. An example is hadrosaur tracks found in late Cretaceous formations in the Canadian province of Alberta. Hadrosaurs were large, bipedal herbivorous dinosaurs. The Alberta footprints show that, like elephants of today, these animals had evolved pads on the soles and digits of their hind feet to cushion their great weight.

Tracks and Behavior

Fossil footprints have proved to be both informative and corrective on the individual and group behavior of the animals that made them. For example, the dinosaurs of the sauropod group, first described in 1842 by the British zoologist Richard Owen, include the largest land animals that ever lived. For more than a century it was assumed these massive quadrupedal herbivores, with their long neck and long tail, must have been habitually aquatic, depending on their buoyancy in water to support their enormous weight. In fact, it was seriously questioned whether they could move about on land at all. Then Roland T. Bird of the American Museum of Natural History discovered sauropod tracks in two Texas formations of Cretaceous age: at Paluxy Creek in 1944 and at West Verde Creek in 1954. His finds decisively settled the issue. A brontosaur trackway from Paluxy Creek, now displayed in the American Museum, is particularly dramatic. It is a narrow trackway, with huge footprints deeply impressed in what could only have been a mud flat either partially or completely exposed to the air. The animal's steps, one fully two meters from the next, are unmistakably those of an efficient terrestrial walker. So much for the assumed habitual aquatic habitat. The drama is supplied by something else. Superposed on the brontosaur tracks are the footprints of a large carnivorous

bipedal dinosaur—possibly *Allosaurus*—moving in the same direction. It seems likely that the carnivore was pursuing the herbivore.

Bird's other site, West Verde Creek, shows in addition to sauropod footprints a continuous deep groove. It evidently was made by the dinosaur's dragging tail, an indication that the mud flat was completely exposed and additional confirmation that sauropods could walk entirely unsupported by water. This does not mean, however, that these giant dinosaurs never went swimming. Another of Bird's findings is informative on the point. This particular sauropod track consists solely of the impressions of forefeet except for a single print of a left hind foot. Apparently the animal's hindquarters and tail were afloat in water shallow enough to enable it to walk along the bottom with its forefeet, just as the hippopotamus does today. Then, when turning left, the sauropod made the extra track as it kicked the bottom with one hind foot.

In depictions of these giant herbivores it has been traditional to show them afloat, contentedly browsing on aquatic plants while their natural enemies, the carnivorous dinosaurs, stand in frustration on the shore, unable to pursue their prey. Tracks in the State Dinosaur Park at Rocky Hill, Conn., suggest a different scene. There hundreds of footprints, most of them of the predatory genus *Eubrontes,* are preserved in lake sediments of early Jurassic age. Some prints are perfectly clear, others are faint; some are deeply imprinted, others are shallow. All show the three digits characteristic of a bipedal carnivorous dinosaur. One trackway in particular, discovered by Walter P. Coombs, Jr., of Western New England College, sheds light on the swimming ability of carnivorous dinosaurs. It consists of eight successive footprints averaging just over one meter between steps. The step with the right foot forward is consistently about 20 centimeters longer than the step with the left foot, which suggests to Coombs that the swimming movements had a kind of galloping rhythm. Further, some of the prints show exceptionally clear claw marks, precluding the possibility that they are ghost prints, and yet give no indication that the animal was putting any weight on its feet. According to Coombs's interpretation, the animal was swimming in shallow water, kicking the bottom with the tips of its toes. Moreover, the track sequence begins and ends abruptly, indicating that the swimming motion intermittently lifted the animal clear of the bottom. Coombs concludes that the traditional image of the frustrated carnivore standing at the water's edge is in need of revision.

Fossil evidence that dinosaurs may have lived in social groups has until recently been quite limited. One example is an accumulation of more than 20 dinosaur skeletons in an early Cretaceous formation at Bernissart in Belgium. All the animals were large bipedal herbivores of the genus *Iguanodon.* It has long been suggested they had died when a whole herd tumbled into a limestone ravine, mute evidence that the iguanodons traveled in herds. Recently David Norman of the University of Oxford has demonstrated that the animals were not all found at the same level in the formation and that the formation is actually a series of sediments. His findings cast serious doubt on the herd interpretation.

Other Social Dinosaurs

A second example is an accumulation of dinosaur skeletons found in a Triassic formation at Ghost Ranch, N.M., in 1947. All the animals belonged to the same small carnivorous genus *Coelophysis.* Again, some see this as evidence that these carnivores hunted in packs and others view it as an unexplainable freak aggregation. Indeed, by their very nature fossilized skeletal remains cannot be expected to yield direct evidence on an animal's behavior.

One source is the nests made by dinosaurs. The discoveries in Mongolia in the 1920's of dinosaur nests complete with unhatched eggs, symmetrically arranged like bird eggs, was the paleontological sensation of its day. One nest, that of a small quadrupedal dinosaur of the genus *Protoceratops,* provided evidence that, unlike such modern reptiles as the turtle, these Cretaceous animals may have guarded their unhatched eggs. Beside the nest was the fossilized skeleton of a predatory coelurosaur—a small bipedal carnivorous dinosaur—that had apparently met a violent end. Its discoverer assigned the fossil to a new genus, appropriately named *Oviraptor.*

A second nest discovery, made only two years ago in Montana, continues the story of dinosaur sociality. In the Two Medicine formation, a deposit of Cretaceous age, John R. Horner of Montana State University found the skeletons of 18 infant hadrosaurs. The young animals were judged to be one or two weeks old. Their continued association with a nest strongly suggests sibling social behavior and also implies the possibility of parental attendance on the young after their hatching.

What social behavior can be deduced from footprints? For one thing, the first clear evidence for dinosaur herd behavior has come from the same two Texas sites studied by Bird. Some 20 of the sauropod tracks he uncovered were parallel to one another, indicating movement in a common direction, and few of the footprints overlapped. Moreover, the prints differed in size, showing that the herd included both young animals and mature ones.

A few years after Bird's discovery John H. Ostrom of Yale University uncovered 25 sets of dinosaur footprints in an early Cretaceous formation at Lake Eanes in Texas. All were made by bipedal herbivores, probably of the genus *Camptosaurus.* All were moving in the same direction: to the northwest. Ostrom has also found evidence suggestive of hunting-pack behavior among carnivorous dinosaurs preserved in the Triassic sediments of Mount Tom in Massachusetts. There on a bedding-plane surface are at least 19 separate tracks of the dinosaur genus *Eubrontes.* Again, all the animals were moving in the same direction. Evidently they were walking abreast, because none of the trackways overlap.

Thus the evidence of the footprints makes it clear that at least some herbivorous and carnivorous dinosaurs were gregarious. This in itself is an advanced mode of social behavior without close parallel among the reptiles of today. On the basis of Bird's Texas tracks Robert T. Bakker of Johns Hopkins University has even suggested these sauropods traveled in "structured" herds, with the juveniles at the center and the adults forming a protective perimeter. Supplementary evidence indicating that the structured arrangement was deliberate rather than a matter of chance would be most welcome.

A Canadian Bonanza

The inventory of fossil footprints has been greatly enriched in the past few years by urgent work on the middle Cretaceous sediments of the Peace River Canyon in British Columbia. The 17-mile canyon has been known to be rich in tracks ever since a pioneer paleontologist, Charles M. Sternberg of the Royal Ontario Museum, uncovered 400 dinosaur footprints there in 1922. He published his findings in 1930, naming eight new types of footprint, all made by dinosaurs and all but one showing a bipedal gait.

In 1975 construction began on a large dam downstream from the site of Sternberg's discoveries. This spurred efforts by the Natural History Museum of Alberta to rescue at least some of the other fossil footprints presumed to exist in unsurveyed parts of the canyon before its inevitable flooding. Five seasons of work under the leadership of Philip Currie have brought spectacular results.

Although only the lower four miles of the canyon could be examined, more than 100 trackways were found. They included a total of 1,700 dinosaur footprints representative of at least 10 different species, a series of bird footprints and a single turtle trackway. More than 1,000 of the footprints have now been mapped and almost as many have been measured. Casts have been made of

Save a bundle!

On MIT's annual bookclearing. You'll find great bargains on scholarly, professional, and technical books in the humanities and social sciences, architecture and urban studies, science and engineering.

Don't miss this once-a-year opportunity to acquire high quality books at substantial savings. The sale ends June 30th and is offered only in the United States and Canada.

Please send me your sale catalog. (type or print)

Name_____

Address_____

City_____

State_____ Zip_____

28 Carleton Street
Cambridge, MA 02142
(Tel. 617-253-2884)

ASA **THE MIT PRESS**

Free!
Edmund Scientific Catalog

4,000 Unique Items

Astronomy, Microscopy, Biofeedback, Weather, Alternate Energy, Binoculars, Optics, Magnets, Magnifiers, Tools, Unique Lighting, Lab Equipment, and much more. Over 4,000 unique and fascinating products. Send for our FREE, colorful Edmund Scientific Catalog... Today!

Rush me your free catalog!

Name _____
Department _____
Institution _____
Address _____
City _____
State _____ Zip _____

Clip And Mail Coupon Today To:
Edmund Scientific, Dept. 8331 2110
Edscorp Bldg., Barrington, N.J. 08007

some 200 of the prints, and 90 prints have been removed for further study and eventual public display.

Among the splendid dinosaur prints is an entire growth series—from juvenile to adult—of an early hadrosaur. The series shows how the shape of the animals' feet changed as they matured and their weight increased, a record unprecedented in ichnology. A herd of these herbivores appears to have traveled across the Cretaceous sediments, its members spread out along a broad front and sometimes walking side by side. Judging by the frequent overprinting of juvenile footprints on adult ones the younger animals followed along after the adults. One series of tracks shows that 11 animals proceeded southward for a time and then suddenly switched to travel eastward. (The preservation of a single hadrosaur trackway moving in the opposite direction, from south to north, shows that the left turn of the others was not dictated by some natural obstruction.) A close study of four of the 11 trackways has even revealed that when one of the hadrosaurs lurched sideways, three of the others veered away to avoid a collision.

The Peace River prints have also added to the evidence that carnivorous dinosaurs, small to medium in size, hunted in packs. For example, at one location six trackways of medium-size carnivores are all headed in the same general direction, although two of the tracks cross each other twice. In contrast the trackways of the larger carnivores—probably tyrannosaurs—found in the

TWO TEXAS TRACKWAYS record a dramatic Mesozoic episode. Discovered by Ronald T. Bird at Paluxy Creek in 1944, the set of larger footprints are those of a quadrupedal herbivorous dinosaur, probably one of the brontosaurs. The smaller three-toed footprints visible to the left of the main track were made by a bipedal carnivorous dinosaur, possibly of the genus *Allosaurus*. The carnivore was moving in the same direction as the herbivore, probably in pursuit.

canyon are those of individuals or at the most pairs. Thus there is still no evidence from any footprint locality in the world to suggest that the larger carnivorous dinosaurs hunted in packs.

Footprints can also yield evidence on how fast the animal that made them was traveling, although care in interpretation is called for. A case in point is provided by dinosaur tracks discovered early in the 1960's at an English quarry. There, near Herston in Dorset, the Purbeck Beds, a late Cretaceous formation, lie exposed. On the quarry floor were found two parallel lines of footprints of a bipedal herbivore. The initial interpretation was that these prints had been made by the opposed feet of a single animal. The modest length of the stride suggested the dinosaur must have moved quite ponderously; one analyst even proposed that since the animal was clearly putting its weight on the ball of its foot, it had been ascending what was then a steep incline at a slow walk. This picture fitted the then current image of dinosaurs as sluggish, slow-moving animals. When the quarrying was extended, however, it was discovered that the two tracks diverged; they had been made not by one dinosaur climbing slowly uphill but by two much faster-moving dinosaurs that were possibly running side by side.

Evidence from the Peace River tracks makes it clear that some dinosaurs were able to move quite rapidly. With a formula developed by R. McNeill Alexander of the University of Leeds the speed of the animal's movement can be calculated from the observed pattern and spacing of its footprints. The highest speeds recorded by the Peace River tracks are those of medium-size carnivores: 16.5 kilometers per hour (10.3 miles per hour). This is close to the highest running speed that can be attained by a human being. The other carnivores traveled at lower speeds: between 6 and 8.5 k.p.h. (3.8 and 5.3 m.p.h.). The herbivores moved more slowly. Their top speed was the same 6 k.p.h. that was the slowest pace of the carnivores.

Overall, then, recent footprint studies demonstrate that the long-prevailing picture of dinosaurs as essentially solitary, slow-moving and even clumsy and slow-witted animals is far from true. It appears likely that when equally careful analyses are done of the fossil tracks of other terrestrial vertebrates—for example amphibians, other kinds of reptiles and the mammals of the early Cenozoic—new information about their behavior will be forthcoming and will prove to be equally at variance with past conclusions. Are we at the beginning of a renaissance in the study of fossil footprints? Certainly it seems safe to predict that over the next few decades this area of investigation will be one of the most exciting in paleontology.

Which of these languages would you like to speak?

Mark the one you want to speak in 2 or 3 months' time

American English	☐	German	☐	Norwegian	☐
Arab	☐	Greek (Modern)	☐	Polish	☐
Chinese	☐	Hebrew (Modern)	☐	Portuguese	☐
Danish	☐	Irish	☐	Russian	☐
Dutch	☐	Italian	☐	Spanish	☐
French	☐	Japanese	☐	Swedish	☐

A Linguaphone Course makes you feel at home in almost any country. You'll never miss the meaning of conservations or be at a loss for words.

- It must work—over 4 million Linguaphone students in 88 countries speak a second language FLUENTLY.
- Proven learning success. You LISTEN to real conversations on cassettes...UNDERSTAND what you hear by following illustrated textbooks. HOLD CONVERSATIONS with the speakers.
- You start speaking the very first lesson.
- You learn at your own convenience.
- You develop an authentic accent. Only native-born speakers are used.
- SAVE MONEY! Comprehensive course—basic and intermediate instruction. ALL IN ONE COURSE.
- Course includes 30 lessons on cassettes...three textbooks...elegant carrying case.
- Order your Linguaphone course today. Talk the language...feel at home in far away places.
- MONEY-BACK GUARANTEE: The Linguaphone Course is unconditionally guaranteed. Try it four weeks, if you're not satisfied, return it for a full refund. There's no risk and no obligation, so order today!

Linguaphone — The Language Masters

NO RISK GUARANTEE — 28 DAY FREE TRIAL

World Language Courses, Inc. Dept. 120 (512) 687-6761 313 Nolana Ave. McAllen, Texas 78501

YES, I accept your risk-free offer to learn a new language. Please send me the course(s) I have checked at only $105.00 per course postpaid.

☐ Check enclosed Charge my ☐ Visa ☐ Mastercard

Acct. No. _____ Expires _____

Name (please print) _____

Street _____

City _____ State _____ Zip _____

Signature _____

Garden Camera CALCULATORS

Hewlett Packard

HP 67	$288.50
HP 97	558.50
HP 32E	48.50
HP 37E	57.95
HP 33C	63.50
HP 34C	74.50
HP 38C	74.50
HP 41C	148.50
HP 41CV	207.50
Accessories	Call
Printer	Call
Card Reader	Call
HP 11C	74.50
HP 12C	111.50
HP 15C	99.50
HP 16C	114.50
HP 10C	59.50

Texas Instruments

TI 58C	78.95
TI 59	168.50
PC 100C	149.50
TI MBA	49.50
Criterion Telescope	499.00

master charge VISA

Prices subject to change without notice

Call for low prices on Nikon, Minolta, Olympus and all Major Brand Cameras

Call 1 (800) 223-0595

Or Send postage and handling to

GARDEN CAMERA

135 West 29 Street, N.Y., N.Y. 10001
Or Visit Our Store: 345 Seventh Avenue
New York, Alaska & Hawaii Call:
Tel: (212) 868-1420—Open Weekdays 8:30–6:00
OPEN SUNDAYS 10–4 p.m. Closed Saturdays

Discwasher Wants to Make You a Video Authority

For your copy of "THE VIDEO REFERENCE", complete the coupon below and return it to Discwasher, along with a check or money order for $1 to cover postage and handling.

Please send me ____ copies of "THE VIDEO REFERENCE". Enclosed is my check or money order for $1/copy to cover postage and handling.

Name

Address

City State Zip

discwasher

1407 NORTH PROVIDENCE ROAD, P.O. BOX 6021, DEPT. SA, COLUMBIA, MO 65205 USA

A DIVISION OF JENSEN an ESMARK Company

NMR Spectroscopy of Living Cells

The chemical reactions of metabolism have traditionally been studied one by one in the test tube. With nuclear-magnetic-resonance (NMR) spectroscopy reactions can be followed as they occur in living tissue

by R. G. Shulman

In the middle of the 19th century Louis Pasteur showed that living cells conduct specific chemical reactions: he demonstrated that yeast, which ferments carbohydrates into substances such as carbon dioxide and ethyl alcohol, consists of living cells. Then at the end of the century the brothers Eduard and Hans Buchner ground up yeast cells and showed that the carbohydrates they added to the resulting solution were turned into the same end products. In this way they established that substances inside yeast cells catalyze chemical reactions. The discovery set a strategy for the advance of biochemistry. Throughout this century biochemists have tried to understand how substances extracted from cells and purified could account for the chemical transformations of which a cell is capable. Different investigators have traced different reactions, clearing paths through the welter of events that make up intracellular metabolism. Although some aspects of metabolism remain unexplored, the cleared reaction pathways, crossing, joining and circling back on themselves, are beginning to match the complexity of the living cell itself.

To the extent that objections persist about the validity of modern biochemistry, they continue to be about reducing the processes of life to sequences of chemical reactions. "The reactions may take place in the test tube," one hears, "but do things really happen that way inside the living cell? And what happens in multicellular organisms?" Here I shall describe how one technique is beginning to answer these questions by detecting chemical reactions as they occur inside cells, tissues and organisms including man. The technique is nuclear-magnetic-resonance (NMR) spectroscopy. It relies on the fact that atomic nuclei with an odd number of nucleons (protons and neutrons) have an intrinsic magnetism that makes each such nucleus a magnetic dipole: in essence a bar magnet. Such nuclei include the proton (H-1), which is the nucleus of 99.98 percent of all hydrogen atoms occurring in nature, the carbon-13 nucleus (C-13), which is the nucleus of 1.1 percent of all carbon atoms, and the phosphorus-31 nucleus (P-31), which is the nucleus of all phosphorus atoms.

In NMR spectroscopy two fields are applied to cells, to tissue or to parts of a living organism. The first field is a strong magnetic field. It causes the nuclear dipoles (that is, the hydrogen-1, carbon-13 and phosphorus-31 nuclei in the sample) to orient themselves so that the dipole of each nucleus is aligned either with the field or against it. Alignment with the field is a state in which the nucleus stores less energy than it does when it is aligned against the field.

Then the second field is applied. It consists of electromagnetic radiation in the radio-frequency part of the spectrum. For any particular strength of the magnetic field in which the sample is placed there is a particular frequency of the electromagnetic radiation for which each photon, or quantum of the radiation, will carry precisely the right amount of energy to allow the alignment of a certain type of nucleus to "flip." Hence if the magnetic field strength is held constant and the radio frequency is varied (or conversely the radio frequency is held constant and the magnetic field strength is varied), there comes a time when those nuclei resonate: they absorb the radio photons. For example, in a magnetic field with a strength of 84,000 gauss the hydrogen-1 nucleus resonates at a frequency of about 360 megahertz (360 million cycles per second), the phosphorus-31 nucleus at about 146 megahertz and the carbon-13 nucleus at about 90 megahertz.

In cells, tissues and organisms the precise frequency of the resonance depends on a further circumstance: the chemical environment of the nuclei. The frequency has a "chemical shift," that is, it differs from what it would be if the nuclei were isolated, because in cells, tissues and organisms the nuclei are a constituent of atoms and the atoms are constituents of molecules. Consider an NMR spectrum showing the resonances (in the form of peaks, or absorption lines) in a sample of the amino acid glutamate. Although carbon 13 has a natural abundance of only 1.1 percent, the peaks from C-13 in the five carbon positions in the glutamate molecule are clearly apparent, and if the spectrum is obtained by varying the magnetic field strength, they extend over a range of about 200 parts per million (.02 percent of the field strength). Two of the carbon atoms in the glutamate molecule are each bound to two oxygen atoms (COO^-); their peaks are close together. These peaks are separated quite widely, however, from the peak of a carbon atom that is bound to a hydrogen atom and from the peaks of two carbon atoms that are each bound to two hydrogen atoms.

A given peak in an NMR spectrum can be assigned to a carbon atom at a particular position in a given molecule by a variety of techniques, the most direct of which is to measure the spectrum after a particular position has been labeled with carbon 13 so that the number of C-13 atoms at that position in a sample of the substance is far greater than one would predict from the natural abundance of C-13. The absorption peak from carbon at that position will be correspondingly enhanced. The technique shows one of the advantages of employing carbon 13 to study the chemical reactions of metabolism, namely that it is possible to follow the C-13 introduced as a label as it passes through various positions in various molecules.

The phosphorus-31 nucleus has different advantages. Since the P-31 nucleus is 100 percent of all phosphorus nuclei, no labeling technique is required. Moreover, the P-31 nucleus gives strong NMR signals from a small number of important compounds. When one applies NMR spectroscopy to a suspension of the bacterium *Escherichia coli,* one typically detects peaks from the three phosphorus atoms in adenosine triphosphate (ATP), a molecule in

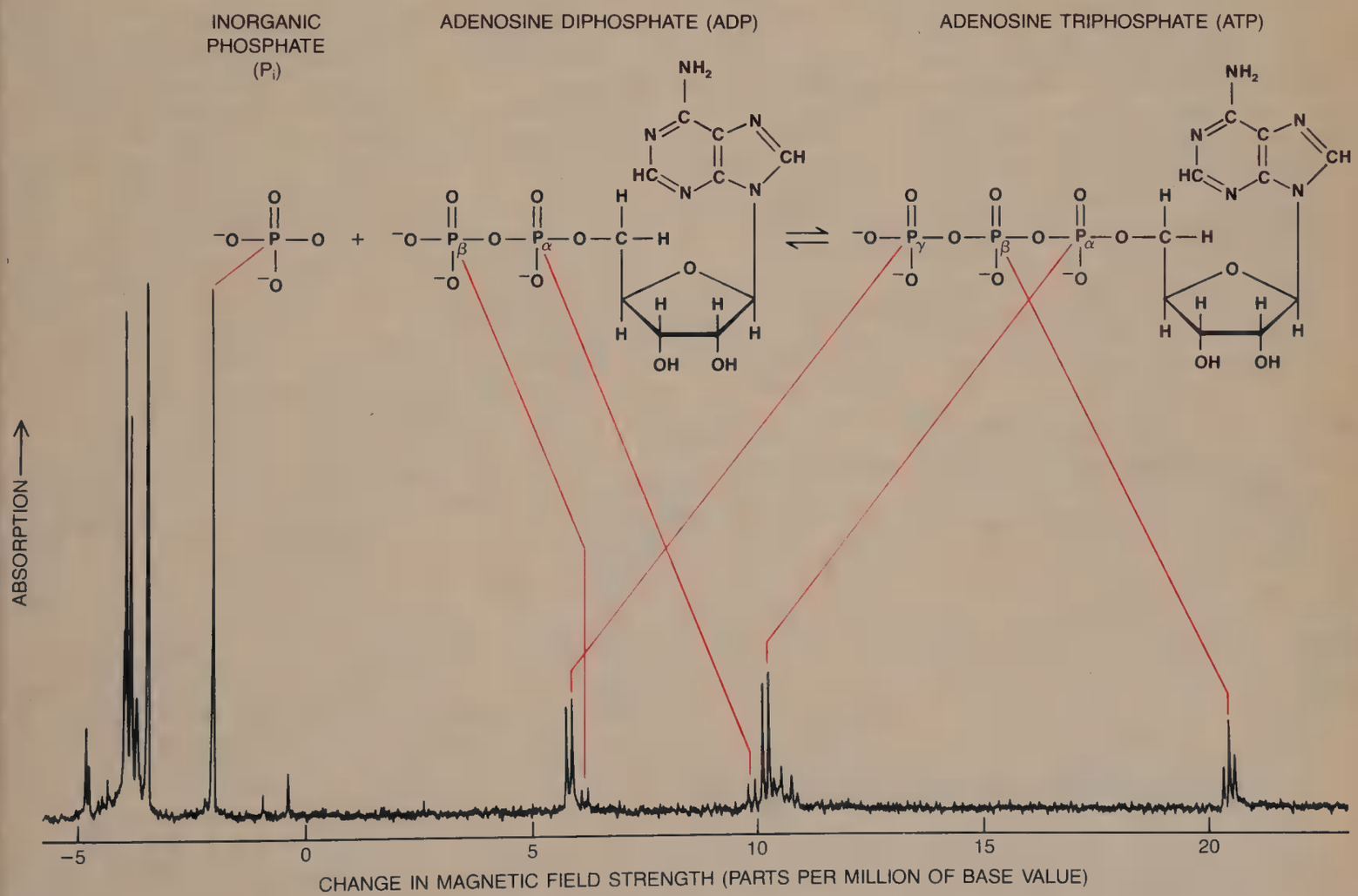
which the cell stores chemical energy. One detects a peak from the phosphorus atom in inorganic phosphate (P_i), which is broken off ATP when its energy is tapped. One detects peaks from the two phosphorus atoms in nicotinamide adenine dinucleotide, a molecule that takes part in intracellular oxidation. Sometimes one detects peaks from adenosine diphosphate (ADP), which has one less phosphorus atom than ATP; from phosphoenol pyruvate, which the cell can tap for P_i to convert ADP into ATP, and from polyphosphate chains.

When the bacteria are living off a supply of the sugar glucose, they utilize the sequence of metabolic reactions known as glycolysis or the glycolytic pathway. Molecules that arise along the sequence, in particular fructose-1,6-biphosphate (FBP), can then be detected too. Such molecules give sharp NMR peaks because each molecule is rather small (a typical one has less than 1,000 times the mass of a hydrogen atom) and small molecules in solution tumble rapidly (on the order of 10^9 times per second). The rapid tumbling means the chemical bonds in the molecule sweep through all possible angles with respect to the environment of the molecule. In effect the tumbling averages out the angles of the bonds with respect to the environment. In contrast, a large molecule tumbles slowly and a molecule locked in place in a solid does not tumble at all. The absorption peaks from a quantity of such molecules are broadened by contributions from nuclei in molecules oriented at various angles with respect to their environment. Even a great concentration of phosphorus-31 nuclei in such a substance may therefore go undetected in NMR spectra because the absorption peaks are simply too broad.

Among the observable phosphorus-31 peaks the one due to inorganic phosphate has been the simplest to understand and among the most informative. Inorganic phosphate is found inside cells in combination with one or two protons. At an intracellular *p*H of 6.7 the populations of the singly protonated form and the doubly protonated form are equal. At higher *p*H's (that is, in a more alkaline environment) the singly protonated form dominates, and at lower *p*H's (a more acidic environment) the doubly protonated form dominates. Each form has a distinctive chemical shift, but the inorganic-phosphate peak in an NMR spectrum of cells will be found in some intermediate position depending on the average degree of protonation of all the P_i in the cells.

Thus the position of the P_i peak yields an accurate measure of the degree of protonation of the P_i, which in turn yields an accurate measure of the intracellular *p*H. Working with me at Bell Laboratories a few years ago, Gil Navon and Kamil Ugurbil showed by these means that the *p*H inside *E. coli* is kept quite constant near 7.4 when the bacteria have a source of energy such as glucose. In the same NMR spectra the conversion of glucose into FBP can be followed, and so can the subsequent increase in the concentration of ATP as the cell stores the energy it has acquired.

The British biochemist Peter Mitchell has proposed that the interconversion of ATP with ADP and inorganic phosphate in the cell is coupled to the passage of protons through the membrane of the cell. Specifically, the reaction that



PEAKS IN AN NMR SPECTRUM represent the absorption of radio waves by atomic nuclei at specific positions in molecules that have been placed in a magnetic field. The top part of the illustration shows the chemical reaction that joins inorganic phosphate (P_i) and adenosine diphosphate (ADP) into adenosine triphosphate (ATP), a molecule in which the living cell stores chemical energy. The bottom part of the illustration shows a phosphorus-31 NMR spectrum of the contents of mouse tumor cells. Almost all the absorption peaks in the spectrum are due to the nuclei of phosphorus-31 (P-31) atoms in the molecules shown at the top. The peak due to the gamma-P-31 in ATP and the peak due to the beta-P-31 in ADP are close together because these phosphorus atoms are in a similar chemical environment. The peak from the beta-P-31 in ATP is more isolated because its environment is unique: it alone is flanked by phosphate groups.

liberates chemical energy by splitting ATP into ADP and P_i serves to expel protons from the cell, thereby creating a proton gradient that stores both electrical and chemical energy. Conversely, the reentry of protons into the cell provides the energy required for the reaction that combines P_i and ADP into ATP. A change in the concentration of protons inside the cell constitutes a change in the intracellular *p*H; hence the chemical shift of the P_i peak in NMR spectra provides a test of Mitchell's hypothesis.

One begins the test by giving glucose to *E. coli* bacteria that have exhausted their food supply. The *p*H in such cells will have dropped to a value considerably less than 7.0. As the cells start to metabolize their fresh supply of glucose, the peak due to inorganic phosphate is observed to split into two peaks. One of them, corresponding to a *p*H of 7.4, arises from P_i inside the cells. The other, corresponding to a lower *p*H (which signifies a greater concentration of protons), arises from P_i in the medium in which the cells are suspended. The peaks confirm that a proton gradient is being established. At the same time the peaks due to ATP are rising.

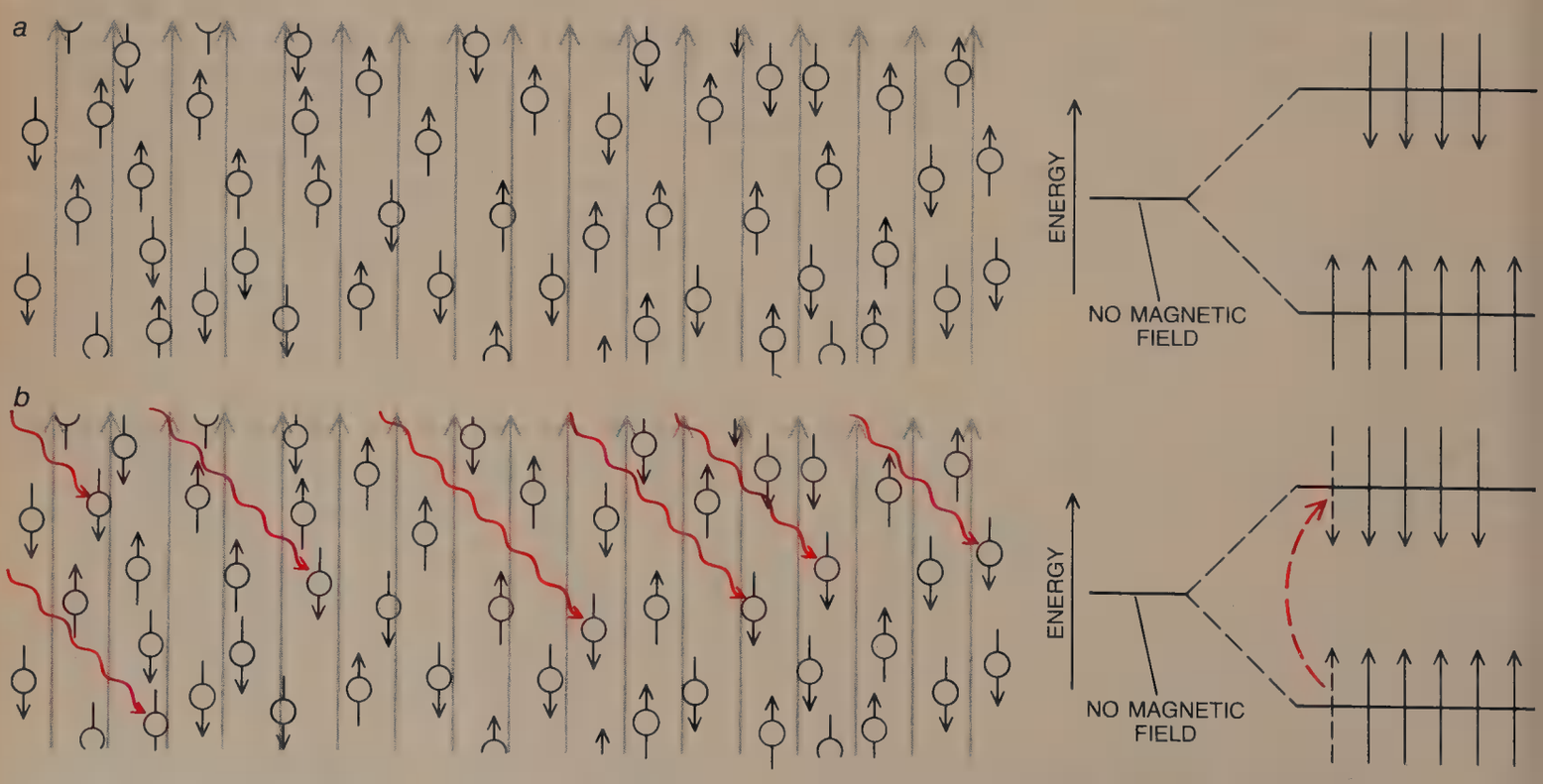
The reaction that splits ATP into ADP and inorganic phosphate is catalyzed by ATPase, an enzyme that couples the reaction to the transport of protons across the cell membrane. If a substance that inhibits the action of ATPase is added to a suspension of *E. coli* along with a fresh supply of glucose, one finds again that the glucose is consumed. Soon FBP and then ATP appear. Yet even when appreciable amounts of ATP have formed, observations of the P_i peak reveal that a proton gradient does not arise. Moreover, the glucose supply lasts much longer than it did before. This shows that an appreciable fraction of the consumption of ATP in the cell goes to maintaining the proton gradient.

The importance of NMR spectra in the study of metabolism is particularly well illustrated by a series of experiments done on muscle by George K. Radda, David G. Gadian and their colleagues at the University of Oxford. In 1974 these workers showed that NMR spectra could be employed to observe metabolism in living muscle tissue from the frog. Their spectra revealed phosphorus-31 peaks from inorganic phosphate, from ATP and from phosphocreatine (PCr), a molecule that muscle cells draw on to convert ADP into ATP. When the muscle was deprived of an external source of energy, the spectra showed a loss of PCr and a constant level of ATP, maintained by the reservoir of PCr. Only when the PCr was exhausted did the level of ATP fall, leaving ADP and P_i. The positions of the ATP peaks revealed that essentially all the ATP in the cells was complexed with positively charged metal ions (thought to be magnesium).

These preliminary studies (supplemented by similar experiments done by Michael Barany, C. Tyler Burt and Thomas Glonek of the University of Illinois Medical Center) succeeded, then, in making measurements that earlier investigators had made by the chemical analysis of numerous muscle samples taken at various times after the tissue had been nourished. Subsequent studies went further. Since the function of muscle is to do work, Gadian, with Douglas Wilkie and Joan Dawson of the University College London School of Medicine, undertook to make the phosphorus-31 peaks in NMR spectra reveal correlations between metabolites and the fatigue of the muscle: the decline of muscle force after prolonged exertion. They suspended a frog muscle in a vessel much like a test tube, which they placed in the NMR spectrometer.

In some experiments the solution surrounding the muscle was oxygenated; in others it was deoxygenated. An electrode stimulated the muscle, causing it to contract. A strain gauge measured the force the muscle developed. Gadian and his colleagues concluded that a decline in the force is well correlated with an increase in the concentration of ADP. Furthermore, it is proportional to the rate at which ATP is consumed. On the other hand, it does not depend on the concentration of ATP. The experimenters could follow the rate of consumption of ATP by measuring the decreasing level of phosphocreatine, the source of ATP in muscle, and simultaneously measuring the level of inorganic phosphate, which arises from the consumption of ATP.

In the past two years advances in NMR spectroscopy and the building of new spectrometers have allowed the



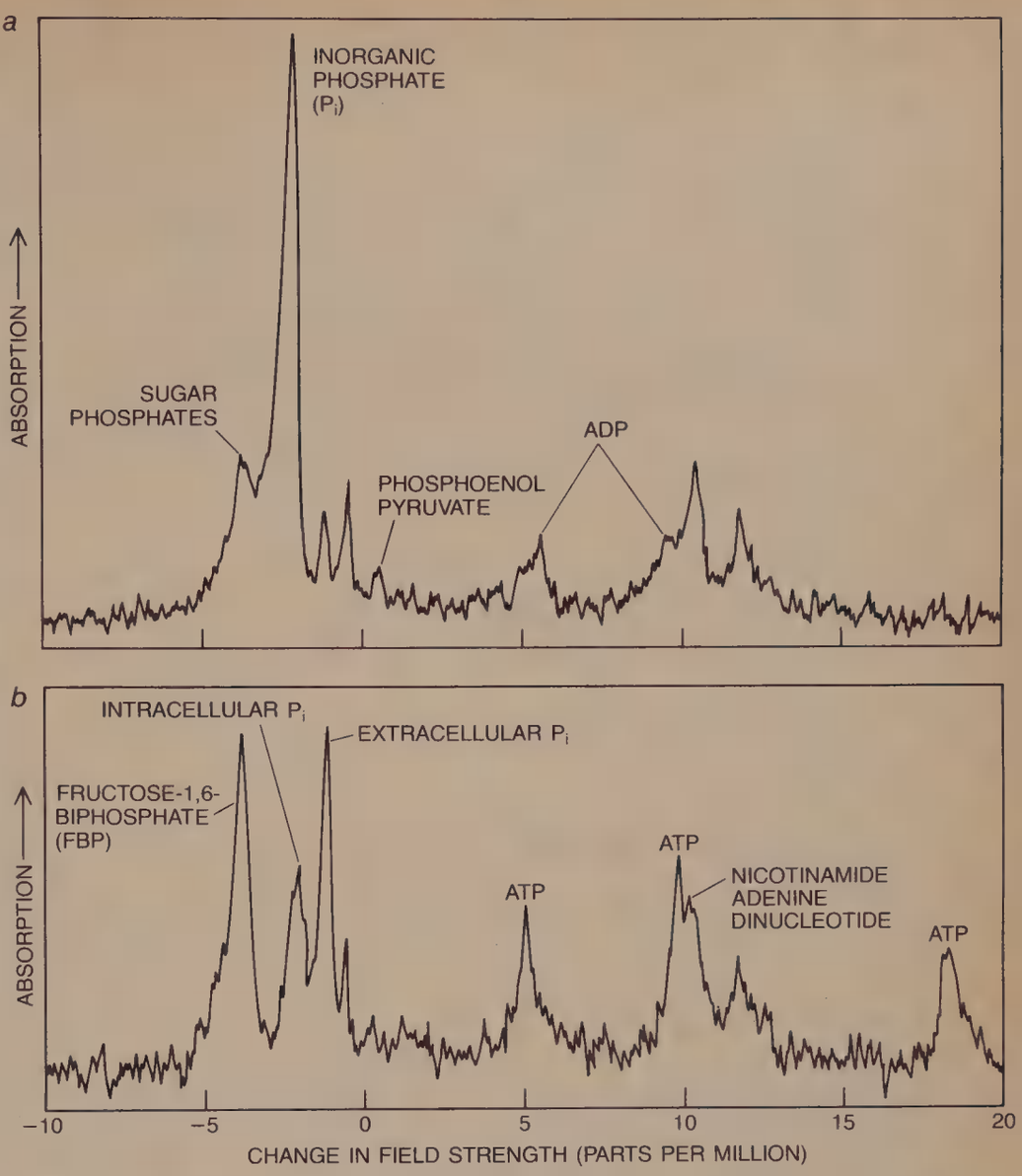
INTRINSIC MAGNETIZATION of certain nuclei (including the nuclei of hydrogen 1, carbon 13 and phosphorus 31) makes NMR spectroscopy possible because each such nucleus is the equivalent of a bar magnet. In a strong magnetic field the nuclei become aligned either with the field or against it (*a*). If the nuclei are then bathed in radio waves, they will absorb the quanta of radiation that give them the energy to "flip" from alignment to antialignment (*b*). The types of nuclei and the chemical environments they are in determine the combinations of radio frequency and magnetic field strength at which absorptions appear as peaks (or "lines") in the NMR spectrum.

study of metabolism in muscle to be extended to the muscles in the limbs of living people. The earlier spectrometers, built for chemical studies, had been designed to accommodate samples in a cylindrical tube. Inside the spectrometer the tube with its contents was surrounded by a coil that emitted radio-frequency radiation, and the tube and the coil were immersed in a static, homogeneous magnetic field. Clearly the examination of intact human limbs would require a different apparatus.

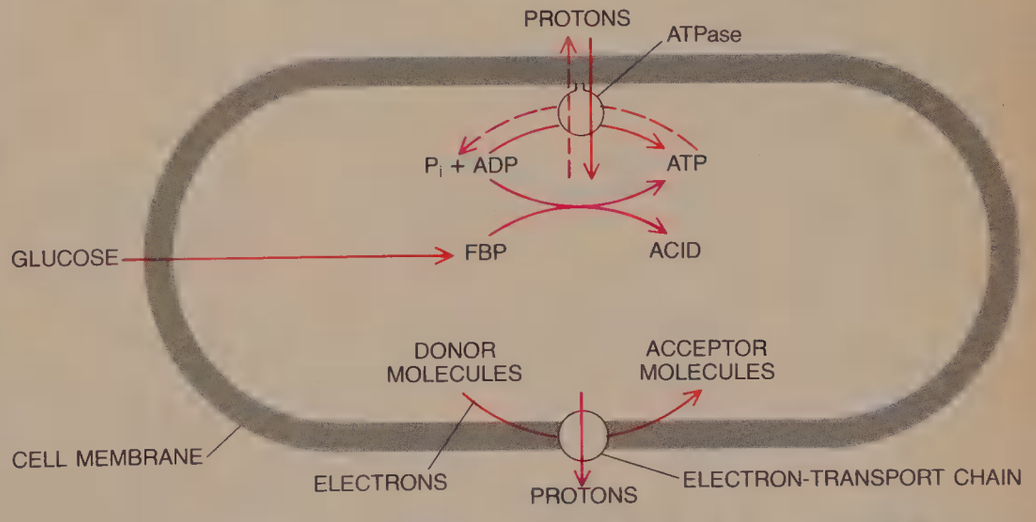
In the new spectrometers a flat coil is laid on the surface of the sample. The flat coil does not excite the entire sample with radio waves; it irradiates only the immediate vicinity of the coil. This turns out to be an advantage: the experimenter can follow chemical reactions in a specific region of tissue. To circumscribe the region still further the magnetic field in the new spectrometers can be adjusted so that it is homogeneous only in a volume of one or two cubic centimeters. Sharp NMR peaks are detected from this volume. In the rest of the volume under the influence of the magnet the magnetic field is quite inhomogeneous, so that if radio waves are absorbed there, the absorption is spread over a wide range of frequencies or magnetic field strengths and merges into the background of the spectrum.

By combining a flat surface coil with a focused magnetic field, Radda and Gadian, in collaboration with workers from Oxford Research Systems, the company that builds the new spectrometers, have been able to measure phosphorus-31 absorption peaks from the liver of living rats to the exclusion of peaks from the surrounding muscle tissue. They knew their spectra represented the liver alone because the spectra showed no peaks from phosphocreatine, which is absent in the liver but is present in muscle in high concentration. More recently Radda, working with a group of clinical investigators led by Brian D. Ross of Oxford, has reported the application of NMR spectroscopy to the diagnosis of human disease. Specifically, the investigators diagnosed McArdle's syndrome in a 51-year-old patient. McArdle's syndrome is a genetic disorder in which the enzyme phosphorylase cannot function. Phosphorylase participates in converting the carbohydrate glycogen into glucose on the demand of muscle cells. Someone who suffers from the syndrome becomes tired after the slightest exertion.

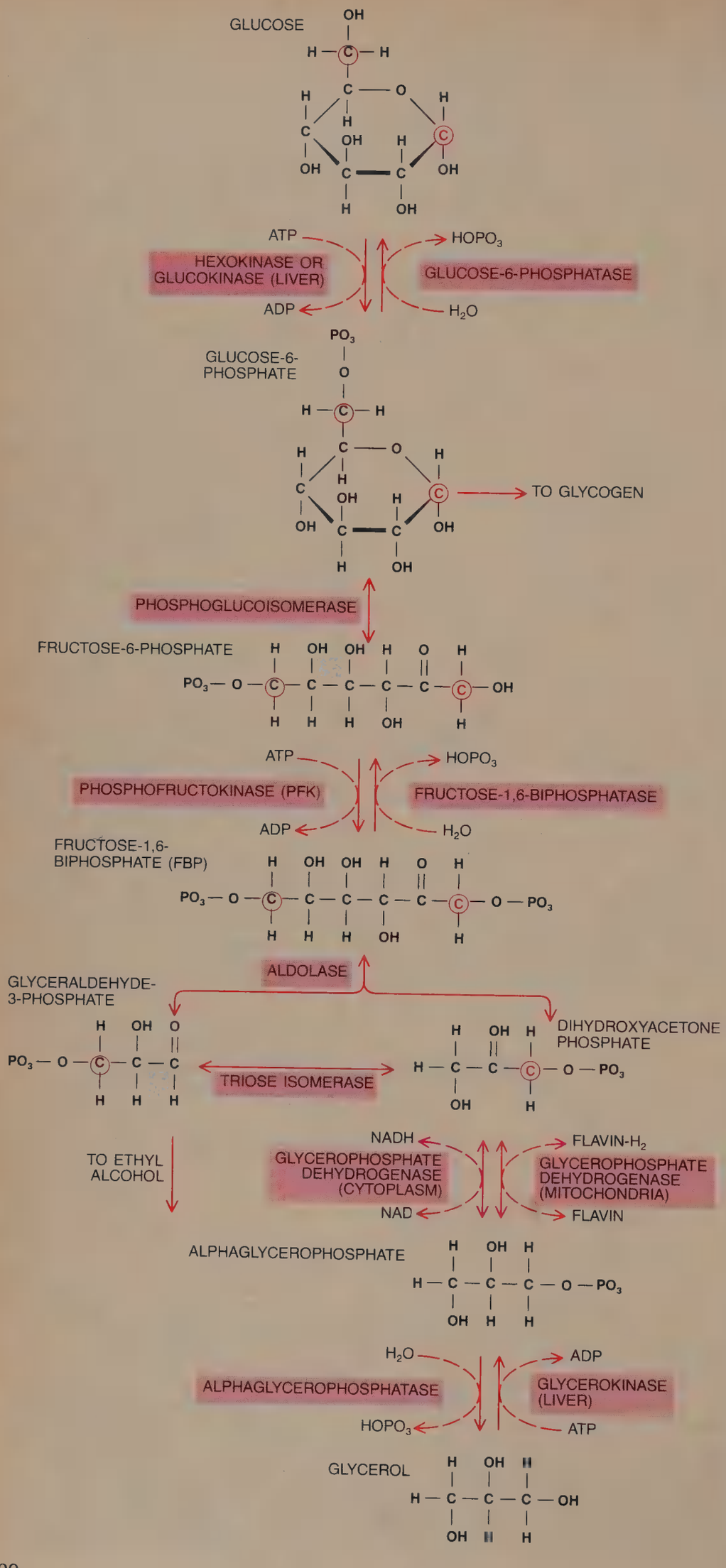
To make the diagnosis the patient was asked to put his arm in the NMR spectrometer, and phosphorus-31 spectra of his forearm were measured as he exercised. The investigators had already shown that in the muscles of normal people the *p*H drops during exercise. In one instance it dropped from 7.0 to 6.4,



EFFECTS OF GLUCOSE CONSUMPTION on the compounds containing phosphorus in the bacterium *Escherichia coli* are traced in these phosphorus-31 NMR spectra, which were made before the cells were given glucose (*a*) and then from four to six minutes after they were given glucose (*b*). In the second spectrum the inorganic-phosphate peak is found to have split. From the positions of the two resulting peaks it can be deduced that the interior of the cells has become more alkaline than the medium in which the cells are suspended. The remaining peaks are due to intracellular substances such as ADP, ATP and fructose-1,6-biphosphate (FBP).



INTRACELLULAR MECHANISMS thought to underlie the changes in the NMR spectra shown in the illustration at the top of the page are diagrammed. The metabolism of glucose gives rise to FBP and ultimately to substances such as lactic acid. In the later steps of the process ADP is converted into ATP. The energy in ATP can be tapped by the membrane-bound enzyme ATPase, which employs the energy to expel protons (hydrogen ions) from the cell. As a result a proton gradient arises and the cell becomes alkaline with respect to its environment. In the reverse reactions, also catalyzed by ATPase, protons are readmitted and the regained energy is utilized to make ATP. The mechanism shown at the bottom expels protons independently.



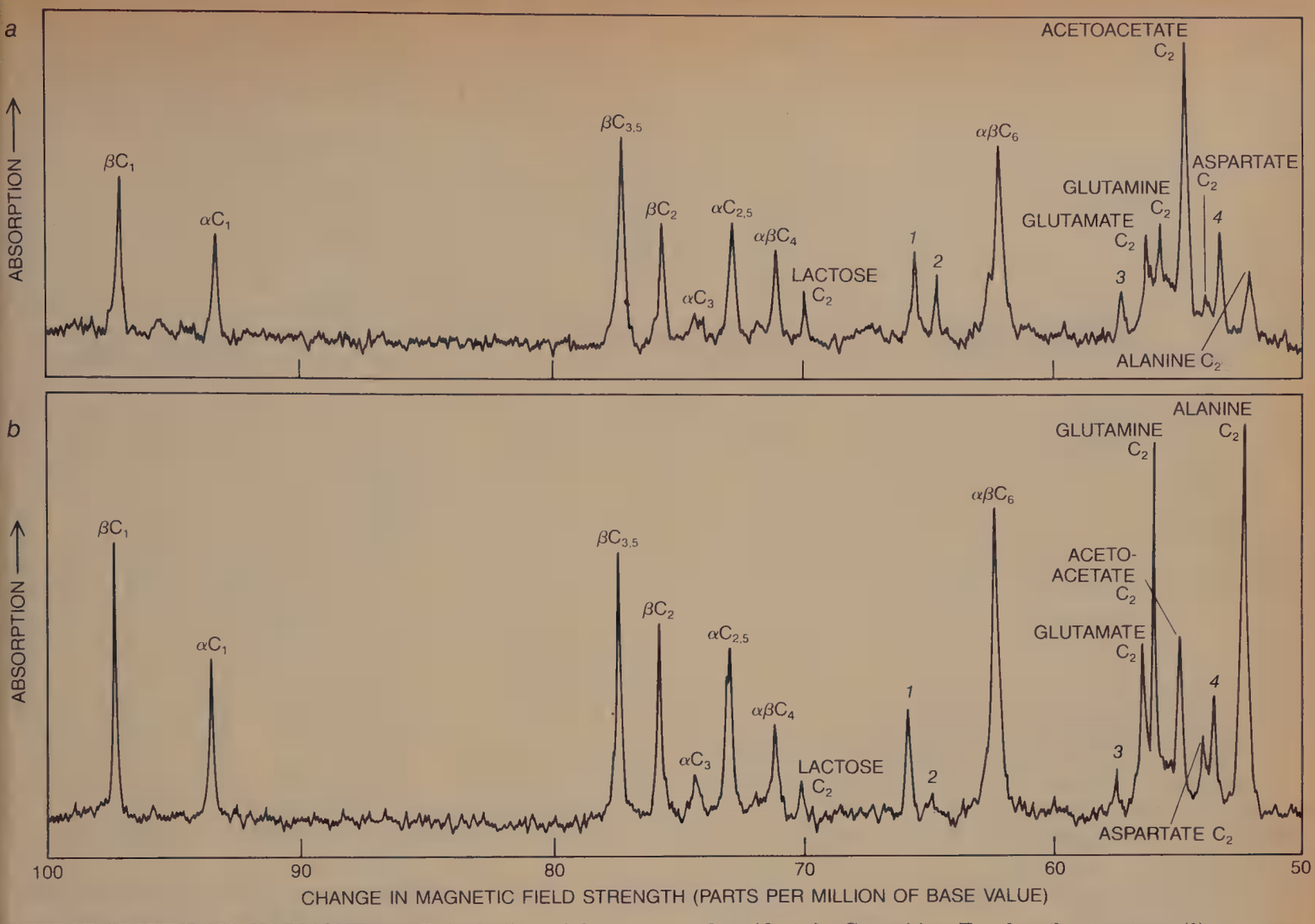
then returned to about 7.0 when the exercising stopped. In contrast, the *p*H of the patient's muscle began at 7.1 and increased to 7.2. It was also observed that the peak in the patient's spectra due to phosphocreatine decreased markedly as the patient exercised.

The explanation of these findings is that in the absence of the action of phosphorylase the patient's muscle cells had little glucose available to be metabolized by glycolysis. The glycolysis would have made lactic acid, whose temporary accumulation lowers the *p*H of normal muscle cells. Glycolysis would also have increased the supply of ATP, which would then have been available to provide the energy needed for muscular exertion. In the patient's muscle cells the required ATP was being made from phosphocreatine. The diagnosis that phosphorylase was inactive in the patient was later confirmed by a chemical test of a muscle biopsy specimen. Ross and his colleagues note that the noninvasive, harmless assays NMR makes possible could be helpful not only in diagnosing diseases such as McArdle's syndrome but also in evaluating the success of treatment.

The details of the glycolytic reactions by which glucose is metabolized can be examined inside the living cell by means of NMR. For example, in the process of glycolysis each glucose molecule, which has six carbon atoms, is converted into two lactic acid molecules, which have three carbon atoms each. As I have noted, the corresponding acidification shows up in phosphorus-31 NMR spectra of healthy muscle tissue as a shift of the inorganic-phosphate peak in accord with the lowered *p*H.

The chemical transformation of glucose can be followed more directly, however, by NMR spectroscopy of cells that are fed glucose in which most of the molecules are labeled with carbon 13 at the carbon position designated C_1. This strategy was pursued in my laboratory at Yale University by J. A. den Hol-

GLYCOLYTIC PATHWAY is the sequence of chemical reactions by which cells metabolize glucose. The sequence can be followed in NMR spectra by feeding cells glucose labeled with carbon 13. Here glucose has been labeled at the carbon position designated C_1 (*colored ball*). Then in the course of glycolysis the label has moved to the C_1 position of FBP and from there into dihydroxyacetone phosphate, a three-carbon compound. If the reverse reactions are occurring, a substantial part of the label will move to glyceraldehyde-3-phosphate (*black and colored ball*) and from there to the C_6 position in FBP. Ultimately it moves to the C_6 position in glucose. The "scrambling" of the label in FBP, measured by the intensity of the carbon-13 peaks in NMR spectra, therefore constitutes a measure of the reversal of glycolysis in the living cell.



EFFECT OF THYROID HORMONE on cells in the liver of the rat is to increase the rate at which they synthesize glucose by a process that is essentially the reverse of glycolysis. One NMR spectrum (*a*) shows peaks due to carbon 13 in substances that rat liver cells have synthesized after the rat was fed alanine (an amino acid) labeled with carbon 13 at the C_3 position. For the other spectrum (*b*) a rat was treated with thyroid hormone and then fed labeled alanine. The peaks due to carbon 13 at various positions in two forms of glucose (α *and* β) are notably higher. Other peaks also show differences. The peaks marked *1, 2, 3* and *4* in each spectrum have not yet been identified.

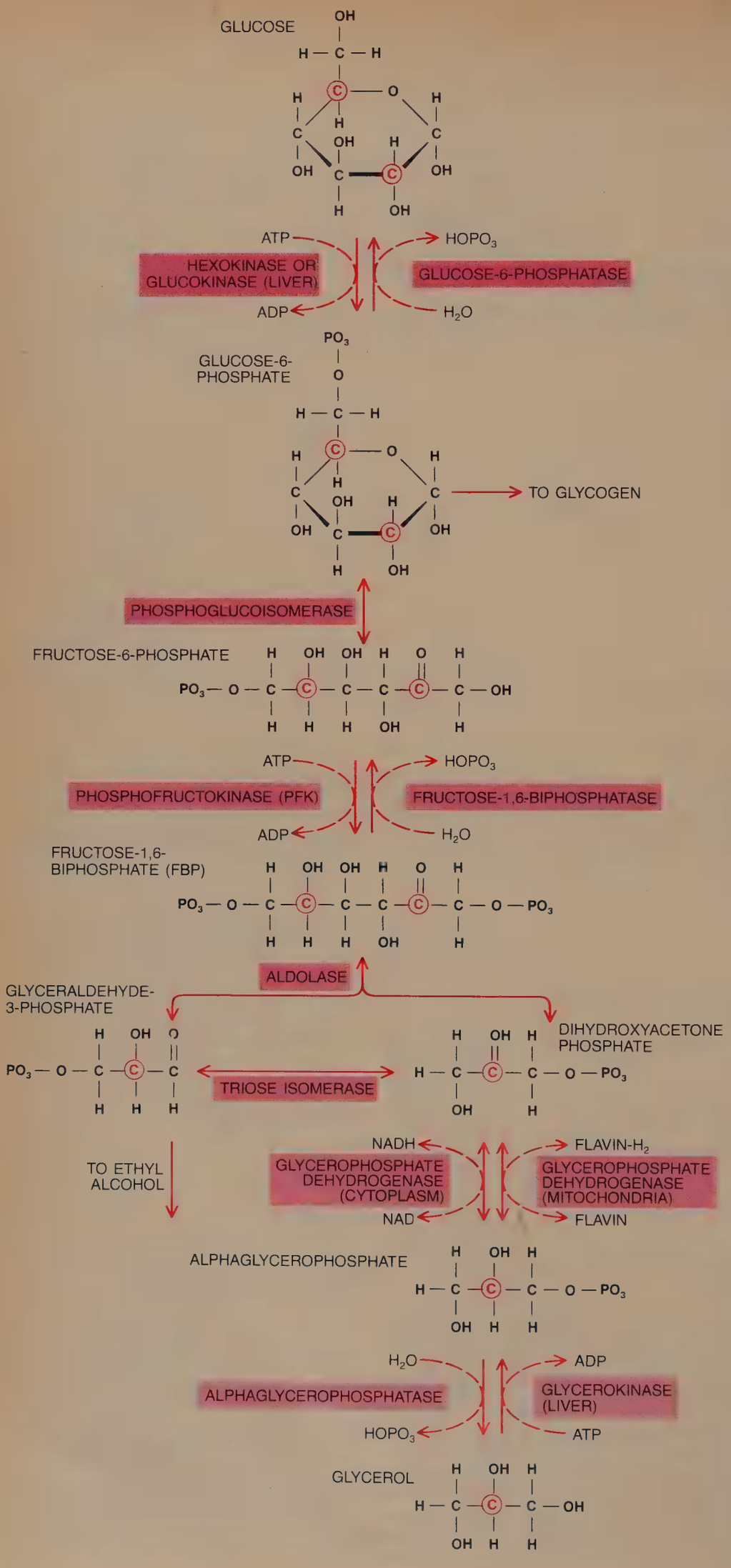
lander, following earlier work done in collaboration with Ugurbil and T. R. Brown. The aim was to measure the rates at which intermediate substances along the glycolytic pathway are transformed from one into the next inside yeast cells, and then to correlate these rates with the properties of the enzymes that catalyze each transformation.

For some time it has been known that the rate at which yeast cells consume glucose is higher in the absence of oxygen than it is if oxygen is present. The increase is called the Pasteur effect. Den Hollander, Ugurbil and Brown began, therefore, by measuring the peaks due to glucose in carbon-13 spectra of a suspension of yeast cells to which they had fed the labeled molecules. The results showed quite clearly that glucose was absorbed and consumed twice as fast in the absence of oxygen.

The next question was whether the Pasteur effect is caused by changes in the rates at which certain enzymes act. One possibility in particular seemed worth examining. Several steps along the glycolytic pathway there is a reaction that transforms fructose-6-phosphate into fructose-1,6-biphosphate, or FBP, under the governance of the enzyme phosphofructokinase (PFK). The enzyme is known to be allosteric: the rate at which it catalyzes the reaction depends not only on the concentration of its substrate (the fructose-6-phosphate on which it acts) but also, at least in the test tube, on the concentration of about 10 different small molecules, which are called for that reason effectors. Partly because of this wealth of possible controlling mechanisms it had long been presumed that PFK is controlled in living cells. No experiment had shown, however, which effectors in a cell really control the enzyme.

Den Hollander, Ugurbil and Brown employed phosphorus-31 spectra to investigate the precise changes caused in yeast cells by the presence of oxygen. Their measurements showed a significant change in the P_i peak. It decreased about fourfold in intensity, and it shifted in position in a way that showed the intracellular *p*H had increased from 7.04 to 7.45. At the same time the peak due to fructose-6-phosphate increased severalfold, and the peaks due to ATP and ADP changed slightly. Test-tube measurements of the activity of PFK made recently by den Hollander and David Reibstein show that under these conditions the enzyme becomes less active.

Remarkably, however, the intracellular concentration of FBP (the product of the reaction catalyzed by PFK) stayed fairly constant in den Hollander's spectra, even though the rate at which FBP formed had markedly decreased. The rate at which FBP was being consumed must also have decreased, and by the same amount. The flux of FBP (the flow of molecular matter through that stage of glycolysis) can be inferred from carbon-13 NMR spectra by a method that capitalizes on the nature of the glycolytic reactions. In the first steps of glycolysis the carbon-13 label den Hollander introduced at the C_1 position of glucose makes its way directly to the C_1 position of FBP. Then comes a reaction in which FBP splits into two three-carbon compounds. Each such compound can undergo reactions that take it further down the pathway. Alternatively, each three-carbon compound can be transformed into the other three-carbon compound and the two can then combine to remake the



FBP. These reverse reactions "scramble" the carbon 13 so that in the remade FBP it appears at the C_6 position.

Overall, therefore, the competition between the downward flow along the glycolytic pathway and the reverse direction of flow, in which FBP reappears, determines the distribution of labeled atoms between the C_1 and the C_6 positions of FBP. If there is little flow down the pathway, there will be ample time for FBP to be remade. In that event the intensity of the NMR peak from carbon 13 at the C_6 position of FBP should equal the intensity of the peak from the C_1 position. If, on the other hand, the downward flow is rapid, there will be little time for the label to be scrambled and the C_6 peak will be nil. Den Hollander's spectra show that the oxygenation of yeast cells increases the scrambling. He was able to calculate that the downward flux through FBP had been reduced severalfold. Hence the flux was being controlled at some point further along the pathway.

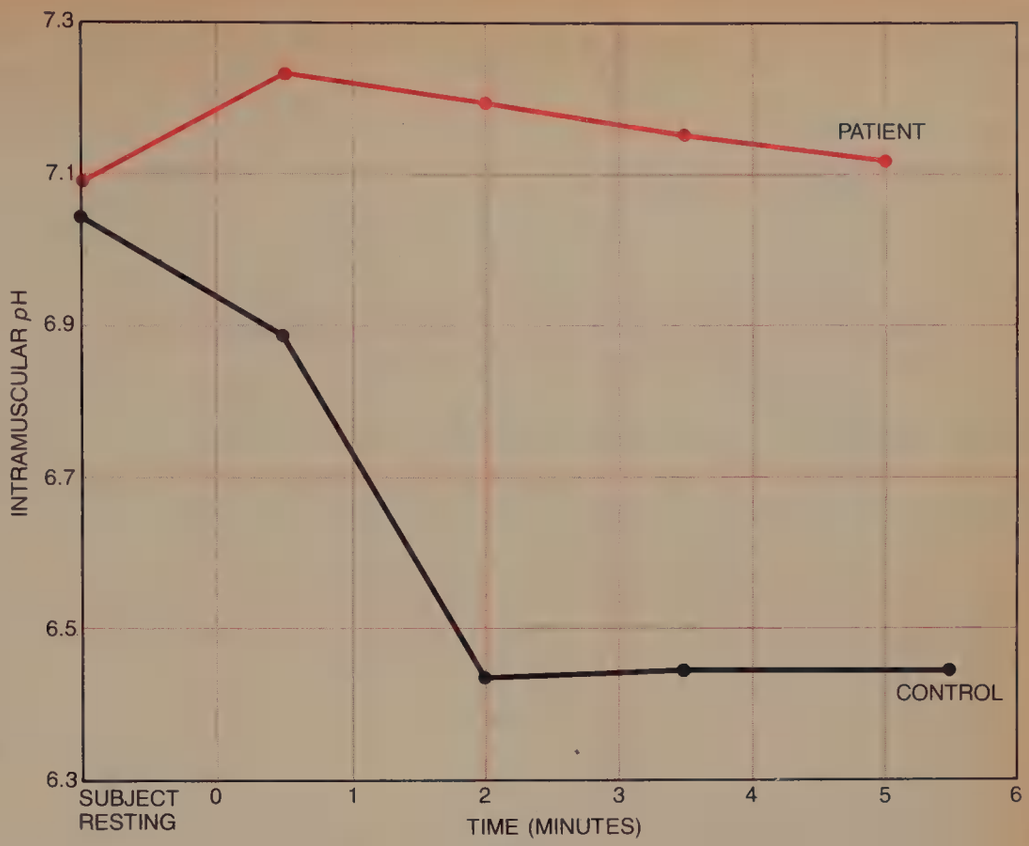
The reversal of flow along the glycolytic pathway is well worth studying because many cells reverse the flow to synthesize glucose if they lack it. FBP is an intermediate in both directions of flow. The enzyme that makes FBP along the pathway leading away from glucose is PFK; the enzyme that catalyzes the opposite reaction, which brings FBP one step closer to glucose, is fructose-1,6-biphosphatase. It is notable that the making of FBP by PFK requires inorganic phosphate to be transferred from ATP to fructose-6-phosphate. In the reverse reaction, however, the phosphate is simply cast off. If both flows occur, the net result is the waste of an ATP molecule in what is called a futile cycle. Many such cycles are known.

What controls a futile cycle? One hypothesis about the cycle involving FBP was stimulated by the knowledge that FBP, like many other sugars, exists in two forms (designated alpha and beta) that differ only in the interchange of two carbon bonds. A mixture of the forms takes only about one second to reach an equilibrium distribution in which 15 percent of the molecules are of the alpha

LABELING OF GLYCEROL (an end product of glucose metabolism) makes it possible to follow the gluconeogenic pathway: the series of reactions that reverses glycolysis. When the middle carbon position of glycerol is labeled with carbon 13 (*color*), the label ends up distributed between the C_2 and the C_5 positions of glucose. A small amount of the label (not shown) ends up at the C_1 position of glucose because it passes through an alternate series of reactions called the pentose-phosphate pathway. NMR studies of gluconeogenesis show that thyroid hormone increases the production of glucose in hyperthyroid rats by increasing the activity of the mitochondrial enzyme glycerophosphate dehydrogenase.

form and 85 percent are the beta. Test-tube studies show that fructose-1,6-biphosphatase, which takes FBP toward glucose, acts preferentially on the alpha form. Aldolase, the enzyme that takes FBP a step further from glucose, acts preferentially on the beta form.

It had accordingly been suggested that the living cell controls the direction of flow through the futile cycle by controlling the distribution of the forms of FBP. The suggestion could not be tested by extracting FBP from the cell and analyzing the distribution. The extraction would take longer than the one second needed for equilibrium to become established. The NMR spectroscopy of intact *E. coli* bacteria makes a different test quite feasible. Den Hollander, Ugurbil and Brown measured carbon-13 peaks from the alpha and beta forms of FBP inside the cells while the cells were metabolizing glucose. The results showed that the FBP was in equilibrium. The direction of flow was not, therefore, being controlled by disequilibrium.



DIAGNOSIS OF DISEASE by means of NMR spectroscopy has been reported by a group of workers led by George K. Radda and Brian D. Ross of the University of Oxford. The disease, McArdle's syndrome, is characterized by fatigue following slight exertion and results from the incapacity of the enzyme phosphorylase, which takes part in making glucose in muscle tissue. The metabolites of glucose decrease the *p*H of cells; in NMR spectra, therefore, the disease is revealed by a lack of change in *p*H, as is shown by the unchanging position of the inorganic-phosphate peak in spectra of the patient's arm made as the patient exercised. The subjects of the measurement wore a cuff to keep fresh glucose from arriving in arterial blood.

When glucose is consumed in the cells of a mammal, the end products of the glycolysis (such as lactic acid) are carried by the blood to the liver along with the end products of the metabolism of fat (such as glycerol). The liver takes these wastes and converts them back into glucose. The liver has thus been made the subject of extensive investigation into the reversal of glycolysis: the gluconeogenic pathway. Paradoxically, the results are too abundant: the study of isolated livers, liver slices and liver cells reveals a number of possible ways the pathway might be controlled but no insight into how it really is controlled in the liver of a living animal. I shall summarize the possible types of control to give a background for discussing the progress being made by NMR spectroscopy in disentangling the actual controls from the merely plausible ones.

First, the rate of production of glucose in the liver can be influenced by the concentrations of the substrates of the gluconeogenic path. An increased level of lactic acid gives rise to a higher rate of glucose production. Second, the rate of production of glucose can be influenced by the supply of food to the organism. Animals deprived of glucose have increased levels of the enzymes that synthesize glucose. Third, the activity of the enzymes that synthesize glucose can be influenced by the extent to which phosphate groups are attached to the enzymes. This phosphorylation, which usually is controlled by rapidly acting hormones such as glucagon, is known to activate a number of enzymes in cells. Fourth, several enzymes that synthesize glucose are controlled by effectors. Fifth, the concentration of various enzymes can be influenced by slow-acting hormones such as the ones synthesized by the thyroid gland.

Sheila Cohen, who was then at Bell Laboratories, employed the NMR spectroscopy of compounds labeled with carbon 13 to follow the gluconeogenic pathway in the intact liver of the rat. One of her studies followed the synthesis of glucose from glycerol, a pathway consisting of six reactions. Several differences emerged between normal rats and rats that had been rendered the equivalent of hyperthyroid by treatment with high doses of a thyroid hormone. In the hyperthyroid rats the rate of production of glucose showed a twofold increase and the concentration of alpha-glycerol phosphate, the compound immediately after glycerol in the pathway, showed a two-and-a-half-fold decrease.

Two enzymes compete for alpha-glycerol phosphate as a substrate. Both are called glycerophosphate dehydrogenase, but one of them is found in the intracellular organelles called mitochondria and the other is in the cytoplasm of the cell. Cohen calculated that under the influence of the hormone the mitochondrial enzyme got a fourfold increase in substrate at the expense of the cytoplasmic enzyme. The finding is in agreement with the levels and activities of the enzymes measured in extracts from liver cells.

There were also some unexpected results. For example, Cohen's NMR spectra showed that the thyroid hormone somehow gave rise to a severalfold increase in the flux through a futile cycle in the gluconeogenic pathway. The basis of such changes is probably the altered concentrations of enzymes brought about over time by the hormone.

A recent experiment done by investigators at Oxford Research Systems and Yale at the instigation of Jeffrey R. Alger in my laboratory suggests the future of NMR spectroscopy. The experimenters fed glucose labeled with carbon 13 to a rat. Then in NMR spectra they watched the glucose disappear from the living animal's stomach. Over a period of hours the level of glycogen in the animal's liver increased; the glucose was being stored as glycogen there. Finally the glycogen peak disappeared as the labeled glycogen was consumed. The NMR spectra had enabled the investigators to follow metabolism throughout the animal's body. Clearly NMR promises the ability to watch precisely how hormones, diet and illness affect the chemical processes of life. Considering the short time in which NMR spectra have served the study of metabolism it is reasonable to expect that within a few years the technique will revolutionize our understanding of how biochemistry proceeds in living animals.

The Physics of Organ Pipes

The majestic sound of a pipe organ is created by the carefully phased interaction of a jet of air blowing across the mouth of each pipe and the column of air resonating inside the pipe

by Neville H. Fletcher and Suszanne Thwaites

No other musical instrument can compare with the pipe organ in power, timbre, dynamic range, complexity of tone and sheer majesty of sound. Like many other musical instruments the organ was brought to a high state of perfection by generations of craftsmen drawing on a slow accumulation of empirical knowledge. By the end of the 17th century the organ had attained essentially its modern form. Two of the 19th century's most illustrious physicists, Hermann von Helmholtz and Lord Rayleigh, reached opposite conclusions about the fundamental mechanism of how organ pipes generate sound but lacked the technical means for resolving the issue. With the advent of oscilloscopes and other modern devices it has become possible to develop a detailed understanding of the mechanism. It turns out that the analyses of Helmholtz and Rayleigh are each valid over certain ranges of the pressure at which air is blown into the organ pipe. The present understanding, which will be described here, is at variance with many of the explanations of organ-pipe mechanisms still found in textbooks.

Pipes cut out of reeds or other hollow-stemmed plants were probably the first musical instruments. They could be made to emit sound by blowing across the end of the pipe, by blowing into the end in such a way that the lips vibrated or by pinching the end in such a way that blowing through it caused the sides of the pipe itself to vibrate. Modern versions of these three primitive wind instruments are flutes, trumpets and clarinets, all of which have been developed so that the performer can play many notes spread over a substantial range of acoustic frequencies.

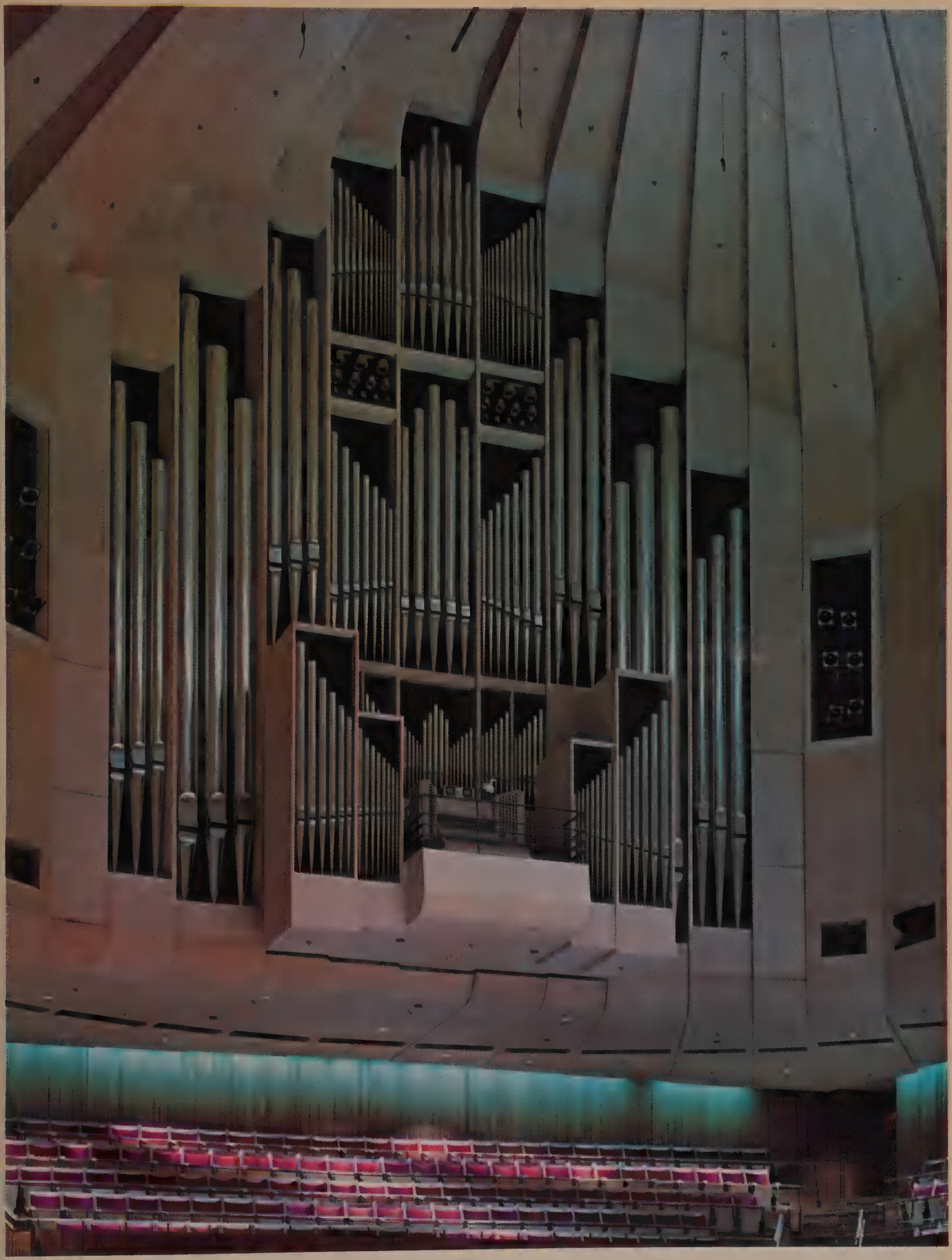
There has been a parallel development in pipe instruments in which a different pipe is dedicated to each note. The simplest such instrument is the panpipe (from "pipes of Pan"), which commonly has about 20 pipes of different lengths, all closed at one end and excited by being blown across the open end. At the other extreme in size and complexity is the pipe organ, an instrument with as many as 10,000 pipes that the player controls through complex mechanical linkages. The pipe organ has a long history. Pottery figures of Alexandrian musicians playing an arrangement of pipes excited by bellows date back to the second century B.C. By the 10th century the organ began to be adopted in Christian churches, and treatises on organ building, written by monks, were in circulation in Europe. The great organ built in the 10th century for Winchester cathedral in England was reputed to have had 400 metal pipes, 26 bellows and two keyboards with a total of 40 keys, each key controlling 10 pipes. In the succeeding centuries there were substantial mechanical and musical advances in organ building, and as early as 1429 an organ with 2,500 pipes was built in Amiens cathedral. By the late 17th century in Germany organs had evolved to their modern form.

The terminology for describing organ pipes reflects their origin in pipes blown by mouth. Organ pipes are open at the top and tapered at the bottom, with a "mouth" (a slot) running across a flattened section above the taper. Inside the pipe is a "languid," or tongue (a horizontal plate), with a "flue" (a narrow slit) between it and the lower "lip" of the mouth. The air that excites the pipe, supplied by a large bellows, reaches the tapered foot of the pipe at a typical pressure of 500 to 1,000 pascals (five to 10 centimeters measured on a water gauge). When the air is admitted to the pipe by the actuation of the appropriate drawstop knob and key, it flows upward and forms a sheetlike jet as it emerges from the flue slit. The jet flows across the mouth and strikes the upper lip, where it interacts both with the upper lip and with the column of air in the body of the pipe to maintain the steady oscillation that generates the pipe's "speech." The problem of how the pipe makes the abrupt transition from silence to steady speech is a difficult one and fascinating in its own right but will not be part of our story. We shall be concerned primarily with the processes that control the steady speech of organ pipes and endow them with their characteristic tone qualities.

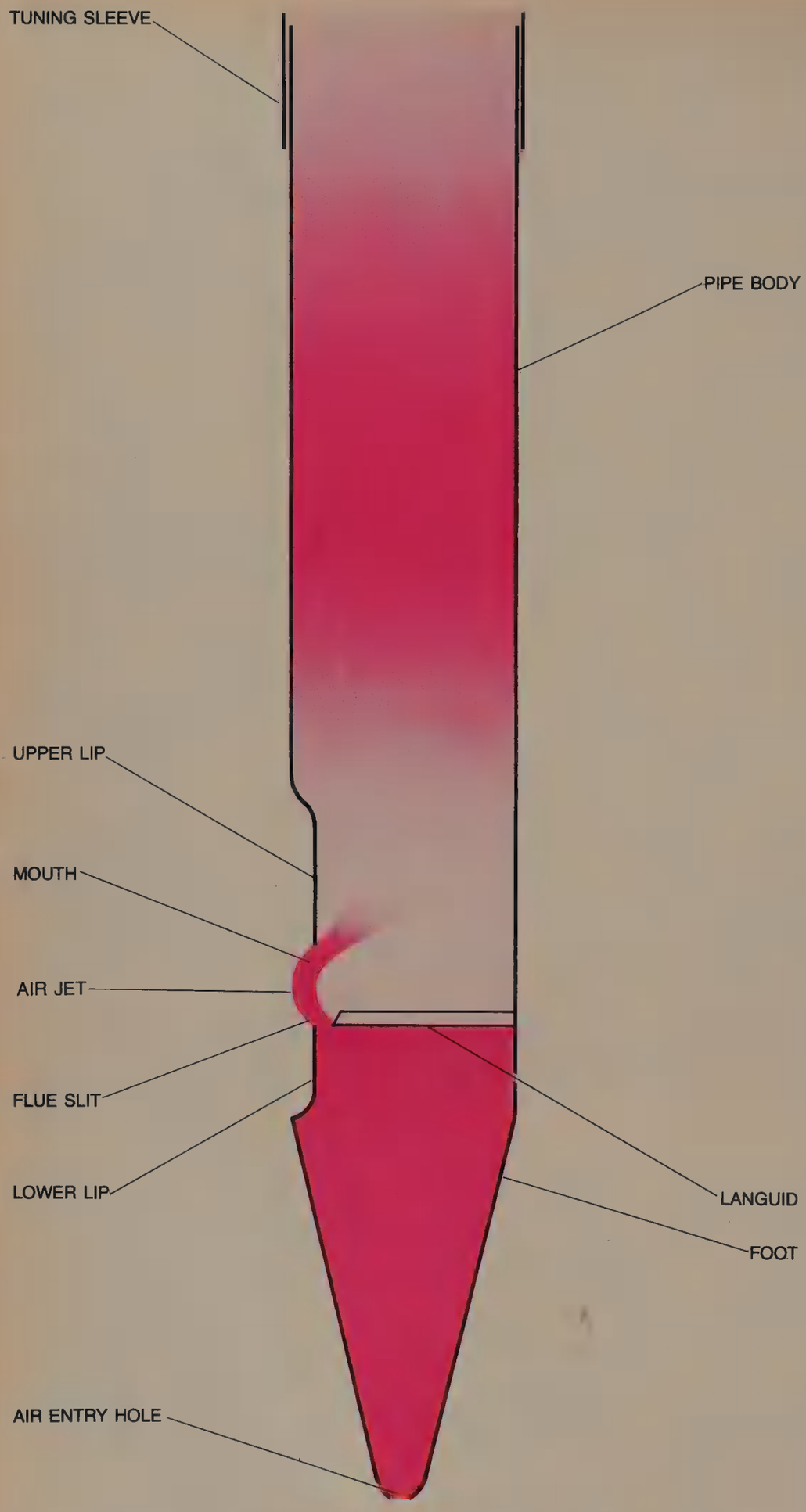
The behavior of the jet that initiates and sustains the organ's speech would seem to present a straightforward problem in fluid flow. It turns out, however, that even a steady jet flowing in a smooth laminar manner is quite difficult to understand theoretically, and the fully turbulent jet that flows in a real organ pipe is impossibly complex. Fortunately the complexity introduced by turbulence actually simplifies the behavior of the jet. If the flow were laminar, the jet would interact with its surroundings through viscosity. In the actual jet viscosity is replaced as the mechanism of interaction by turbulence, and on a scale directly related to the width of the jet. Organ builders go to some lengths to ensure that the jets in their pipes are fully turbulent by cutting fine nicks along the edge of the languid. Paradoxically a turbulent jet is stable and reproducible and a laminar jet is not.

A fully turbulent jet mixes gradually with the surrounding air, thereby broadening and slowing down in a simple manner. When the jet's velocity is plotted against the distance from its central plane, the resulting curve has a bell-like shape, with the maximum velocity in the center. The width of the jet increases linearly with the distance from the flue slit. The momentum of the flow must be conserved, which means that the velocity decreases with the square root of the distance from the slit. This description can be supported theoretically and agrees with experiment (when a small transition region near the flue slit is included).

In an organ pipe that is already excited and emitting sound the jet emerges from the flue slit into the intense sound field in the mouth of the pipe. The air motion associated with the sound is directed into and out of the mouth and therefore at right angles to the plane of the jet. Half a century ago Burniston



CONCERT HALL ORGAN at the Sydney Opera House in Australia, completed in 1979, is one of the largest and most advanced in the world. Designed and built by Ronald Sharp, it has some 10,500 pipes controlled by the mechanical action of five keyboards and a pedal board. The mechanical action, which regulates the flow of air into the pipes, is duplicated by an electric action that is under microprocessor control. The organ can therefore be operated by a magnetic tape on which an original performance has been recorded digitally.



ORGAN PIPE is excited by air blown in from the bottom. The air is formed into a jet by the flue slit between the lower lip and the languid: a plate running across the pipe. When the air blows across the mouth of the pipe, it interacts with the column of air in the pipe at the upper lip, blowing alternately into and out of the pipe. Waves propagating along the turbulent jet maintain a steady oscillation in the air column, causing the pipe to "speak." The standing-wave pattern of pressure in the pipe is represented by the colored shading. The upper end of the pipe is surrounded by an adjustable sleeve, as is shown here, or is closed by an adjustable stopper. Either makes it possible to change the length of the air column slightly for tuning.

Brown of University College London made beautiful photographs of smoke-laden laminar jets emerging into a sound field and found they developed sinuous waves that grew as they traveled along the jet until the jet broke up into a double row of vortex rings rotating in opposite directions. It was the somewhat naive application of these observations and similar ones that led to much of the confused discussion of organ-pipe physics still found in many textbooks.

A more fruitful way to study the behavior of real jets in a sound field is to remove the organ pipe and generate the sound field by means of a loudspeaker. Such studies, conducted by John W. Coltman of the Westinghouse Electric Corporation and by our own group at the University of New England in Australia, have provided much of the current understanding of the physics of organ pipes. Actually Rayleigh had developed a careful and nearly complete mathematical description of the behavior of laminar jets of nonviscous fluids. Because of the fortunate circumstance that turbulence has been found to simplify jet behavior rather than complicate it, Rayleigh's treatment can be carried over with little modification to describe the real jets produced and analyzed experimentally by Coltman and by us.

If one were to forget about the flue slit in the organ pipe, one would expect that the jet, being just a sheet of moving air, would simply be moved back and forth by the acoustic vibrations along with all the other air in the mouth of the pipe. As the jet leaves the flue, however, it is effectively held still by the flue slit itself. This has the same effect as superimposing on the general back-and-forth motion in the sound field an exactly balancing displacement localized at the flue. The localized displacement, which matches the sound field in frequency and amplitude in order to maintain zero displacement at the flue, is carried along with the moving air in the jet and imposes a sinuous wave motion on it.

As Rayleigh showed for a limiting case in his jets, and as we have verified in detail for diverging turbulent jets, the wave motion propagates along the jet at a speed that is a little less than half the speed of the air at the central plane of the jet. In addition, however, the wave grows in amplitude nearly exponentially as it propagates along the jet. In a typical case the amplitude of the wave doubles in traveling about one millimeter along the jet, so that the effects of the wave rapidly dominate the simple back-and-forth lateral motion imposed by the acoustic vibrations.

We find that the growth rate of the wave is greatest when its wavelength along the jet is about six times the width

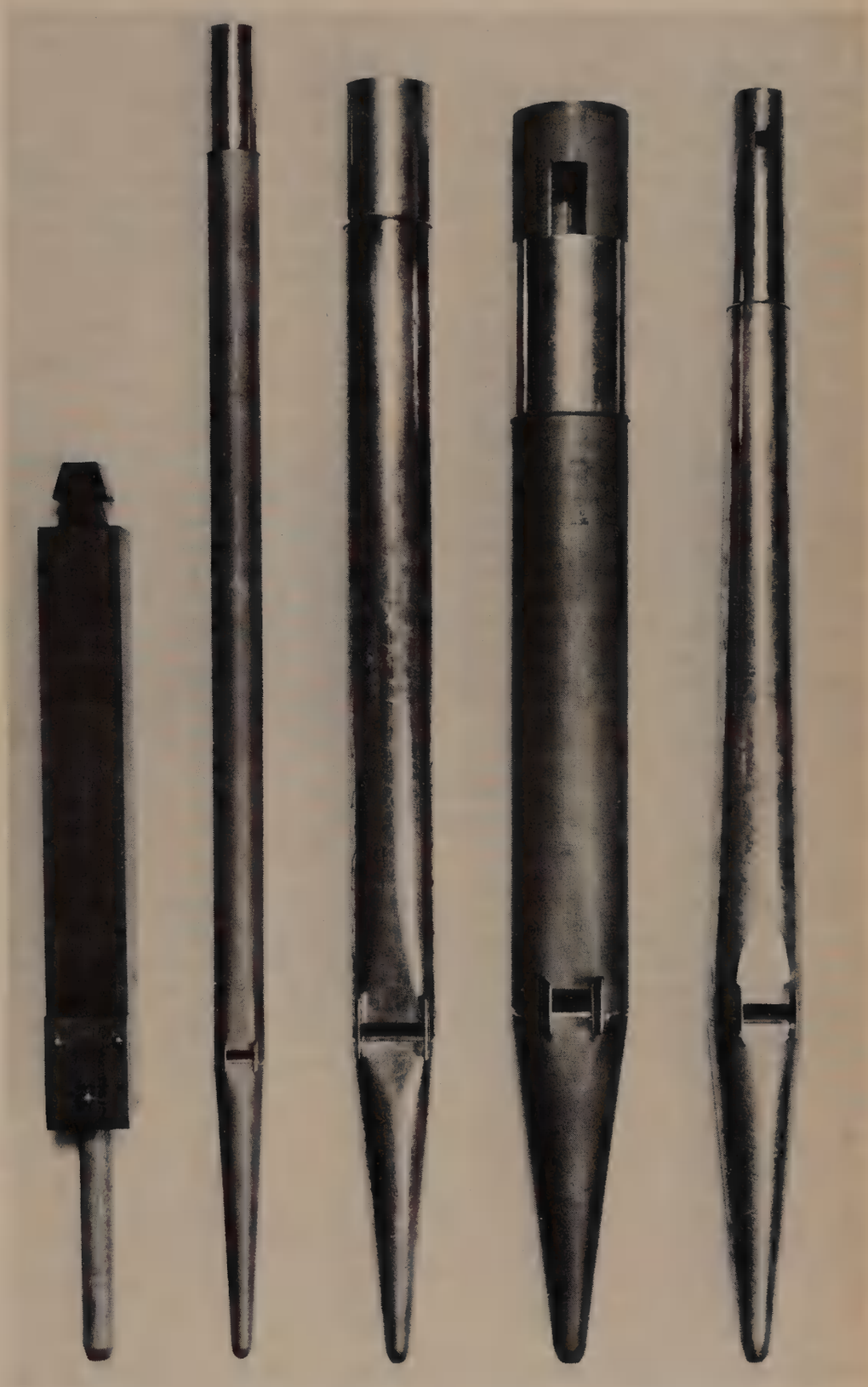
of the jet at that point. On the other hand, when the length of the wave is less than the width of the jet, the wave does not grow at all and may even die out. Since the jet slows and broadens as it moves away from the flue slit, only the long waves, that is, those of low frequency, can propagate with large amplitude along long jets. This fact will be an important one when we come to consider the harmonic content of organ-pipe sounds.

Let us now consider the effects of the organ pipe's sound field on the jet. As one can readily imagine, the large acoustic waves associated with the sound field in the pipe mouth will cause the tip of the jet to flick back and forth across the upper lip of the mouth so that the jet will flow alternately into the pipe and outside it. The situation resembles what happens when one pushes a swing that is already swinging. The air column in the pipe is already oscillating, and if the puffs of air enter the pipe in step with the oscillations, the puffs will maintain the oscillations against sundry energy losses, such as the radiation of sound from the pipe and the drag at the pipe walls. If, on the other hand, the puffs are out of step with the oscillations of the air column, the jet will tend to damp the oscillations and the sound will die away.

The relations between the acoustic motions of the air in the pipe mouth and the time the air pulse arrives inside the upper lip are determined simply by the time required for a wave on the jet to travel from the flue slit to the upper lip. The organ builder calls this distance the cut-up. If the cut-up is large or the blowing pressure is low so that the jet velocity is low, the travel time will be long. Conversely, if the cut-up is small or the blowing pressure is high, the travel time will be short.

In order to discover just what the correct phase relation between the oscillation of the air column in the pipe and the arrival of jet pulses inside the upper lip is, it is necessary to know more about how the pulses act on the air column. Helmholtz thought that the dominant factor in the relation was the volume of flow contributed by the jet. Then if the pulses in the jet flow were to impart as much energy as possible to the oscillation of the air column, they should enter the pipe at the times when the acoustic pressure inside the upper lip was at a maximum.

Rayleigh took a different view. He argued that since the mouth is actually not very far from the open end of the pipe, little acoustic pressure can build up just inside the mouth for the jet flow to work against. He considered that the jet flow was essentially stopped as soon as it entered the pipe, thus quickly building up pressure that could act on the flow in the pipe. Therefore according to Rayleigh the maximum amount of energy would be transferred from the jet if the jet flow entered the pipe when the acoustic flow, not pressure, was at a maximum. The difference between these two maximums amounts to a quarter of an oscillation period in the oscillation frequency of the air column in the pipe. In terms of the swing analogy the difference is between giving the swing a push when it is at the top of its arc and has acquired its maximum potential energy (Helmholtz) and giving the swing a push when



FIVE PIPES OF DIFFERENT DESIGN yield a note of the same pitch but different tone qualities. The second pipe from the left is the dulciana, which has a sweet, thin tone resembling that of a stringed instrument. The third pipe is the open diapason, which has the bright, ringing sound most characteristic of an organ. The fourth has the sound of a very dull flute. The fifth is the *waldflöte* (forest flute), which has a soft sound. The wood pipe at the left is closed with a stopper. It has the same fundamental frequency as the other pipes but only odd harmonics: overtones with frequencies that are odd-numbered multiples of the fundamental. The other pipes differ slightly in length because "end corrections" are needed to give them the same pitch.

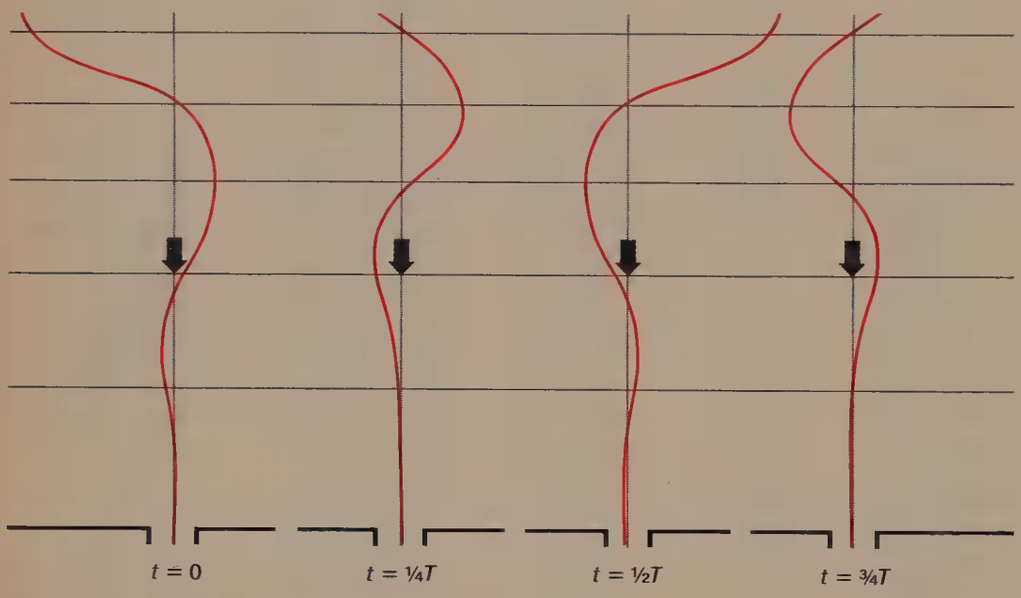
it is at its lowest point and is moving fastest (Rayleigh).

The problem remained unresolved, and indeed virtually uninvestigated, for nearly 80 years. It has now been fairly satisfactorily decided as a result of work by Lothar Cremer and Hartmut Ising of the Heinrich Hertz Institute for Vibration Research in Berlin, by Samuel A. Elder of the United States Naval Academy, by Coltman and by us. Briefly, both Helmholtz and Rayleigh were partly right. The balance between the two driving mechanisms depends on the blowing pressure and the frequency of the sound, with the Helmholtz mechanism dominating for low blowing pressures and high frequencies and the Rayleigh mechanism dominating for high blowing pressures and low frequencies. For an ordinary organ pipe the Helmholtz mechanism is usually more important.

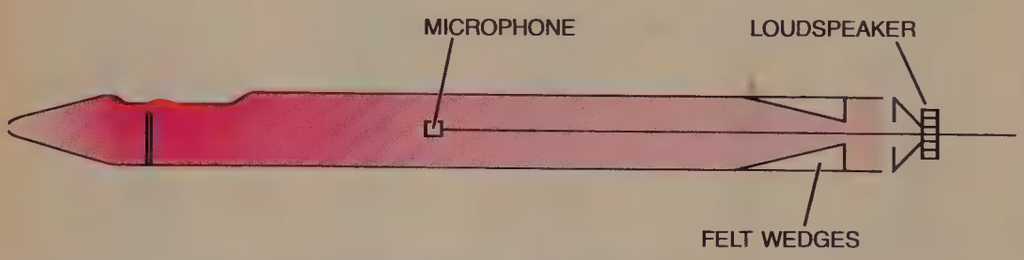
A simple and effective method for studying the properties of the jet was devised by Coltman and has been modified and extended in our laboratory. Essentially one examines the behavior of a jet at the mouth of an organ pipe that has been fitted with wedges of absorbent felt or foam at the far end so that the pipe is prevented from speaking. A sound wave is then directed down the pipe from a loudspeaker outside the far end. The wave is reflected from the mouth end, with and without the jet blowing. In both cases the incident and reflected sound waves interact inside the pipe to create a pattern of standing waves. By measuring with a small probe microphone the changes in the pattern when the jet is turned on, we can determine whether the jet adds energy to the reflected wave or subtracts it.

What our experiments actually determine is the acoustic "admittance" of the jet, which is defined as the ratio of the acoustic flow out of the mouth, produced by the presence of the jet, to the acoustic pressure just inside the mouth. The acoustic admittance has both a magnitude and a phase angle, which can be plotted as a function of frequency or of blowing pressure. When either the frequency or the blowing pressure is varied separately and the admittance is appropriately plotted, it takes the form of a spiral [*see top illustration on opposite page*]. The distance from the origin of the spiral represents the magnitude of the admittance and the angular position along the spiral represents the delay in the phase of the sinuous wave that is imposed on the jet by the acoustic oscillations in the pipe. A delay of one wavelength corresponds to 360 degrees around the spiral. Because of the special properties of turbulent jets it turns out that if the admittance is multiplied by the square root of the blowing pressure, all the measurements for a particular organ pipe fall in the same spiral.

If we keep the blowing pressure constant and increase the frequency of the input sound wave, the point representing the admittance spirals inward toward the center in a clockwise direction. If we hold the frequency constant and increase the blowing pressure, the point spirals outward in the opposite direction.



SHAPE OF THE AIR JET is shown in consecutive "snapshots" as it emerges from the flue slit into the acoustic velocity field generated in the mouth of the pipe by the resonating air column inside the pipe. The periodic displacement of the air in the mouth creates a sinuous wave that propagates with a speed equal to half of the speed of the air at the central plane of the jet and grows exponentially until its amplitude exceeds the width of the jet. The horizontal divisions show distances traveled by the jet wave in successive quarter periods of the oscillation period *T*. The lines get closer together as the jet loses speed. In an organ pipe the upper lip is positioned at about the location of the arrow, so that the air jet blows alternately into and out of the pipe.



ACOUSTIC GENERATING PROPERTIES OF THE JET can be measured by inserting felt or foam wedges in the open end of the pipe to keep it from sounding and introducing a sound wave of small amplitude from a loudspeaker. The sound wave is reflected from the opposite end of the pipe and interacts at the mouth with the air jet. The interaction of the jet with the standing-wave pattern in the pipe is examined with a movable microphone. In this way one can establish whether the jet is adding energy to or subtracting energy from the wave reflected at the base of the pipe. In order for the pipe to sound, the jet must add energy. The measurements are in terms of the acoustic "admittance": the ratio of the acoustic flow out of the mouth to the acoustic pressure just inside the mouth. A plot of the admittance for various combinations of blowing pressure and frequency yields the spiral curve at the top of the opposite page.

When the point representing the admittance lies in the right half of the spiral, the jet is taking energy from the pipe flow and energy is therefore being dissipated. When the point lies in the left half, the jet is imparting energy to the pipe flow, which means that the jet is acting as an acoustic generator. When the point is in the upper half of the spiral, the jet lowers the natural resonance frequency of the pipe; when the point is in the lower half, the jet raises the natural resonance frequency. The reference angle from which the phase delay is measured depends on whether the Helmholtz or the Rayleigh excitation mechanism is dominant, and this, as we have remarked, depends on the blowing pressure and the frequency. It is never far, however, from the zero, or three-o'clock, position on the right arm of the horizontal axis.

Since 360 degrees around the spiral represents a phase delay of one wavelength in the sinuous wave propagated along the jet, it follows that for phase delays from rather less than a quarter of a wavelength to nearly three-quarters of a wavelength a representative point on the spiral will lie to the left of the center line, or in the region where the jet acts as an acoustic generator. We have

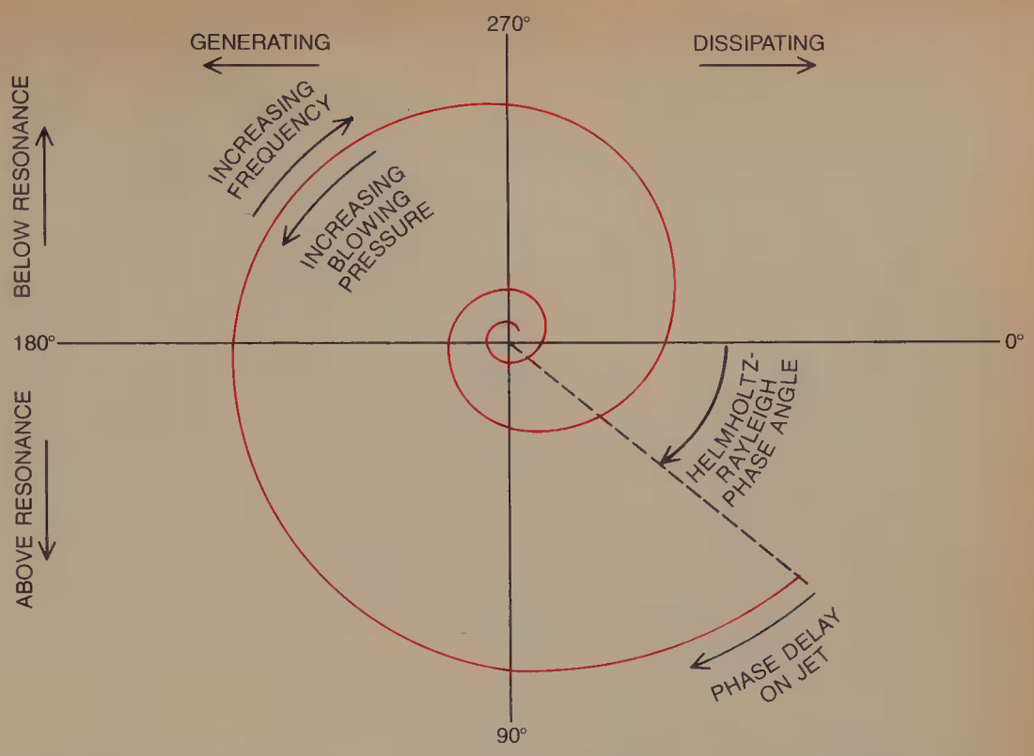
also seen that at a constant frequency the phase delay is a function of blowing pressure, which controls both the speed of the jet and the speed with which the sinuous wave propagates on the jet. Since the wave speed is half the jet speed and the jet speed is proportional to the square root of the blowing pressure, the blowing pressure has to change considerably to alter the phase of the jet wave by as much as half a wavelength. Actually there is about a ninefold range of pressure over which the pipe can emit its fundamental sound if other conditions are right. In practice, however, the pipe will overblow to a mode of higher frequency before the upper end of this pressure range is reached.

We should point out that the spiral can have more than one loop extending far enough into the left side to overcome dissipative losses in the pipe and generate a stable sound. The second such loop, which is generally the only other one that can cause the pipe to sound, corresponds to about three half wavelengths on the jet. Because the jet has only a small admittance at this point the sound generated is weak compared with the sound for a point in the outer loop of the spiral.

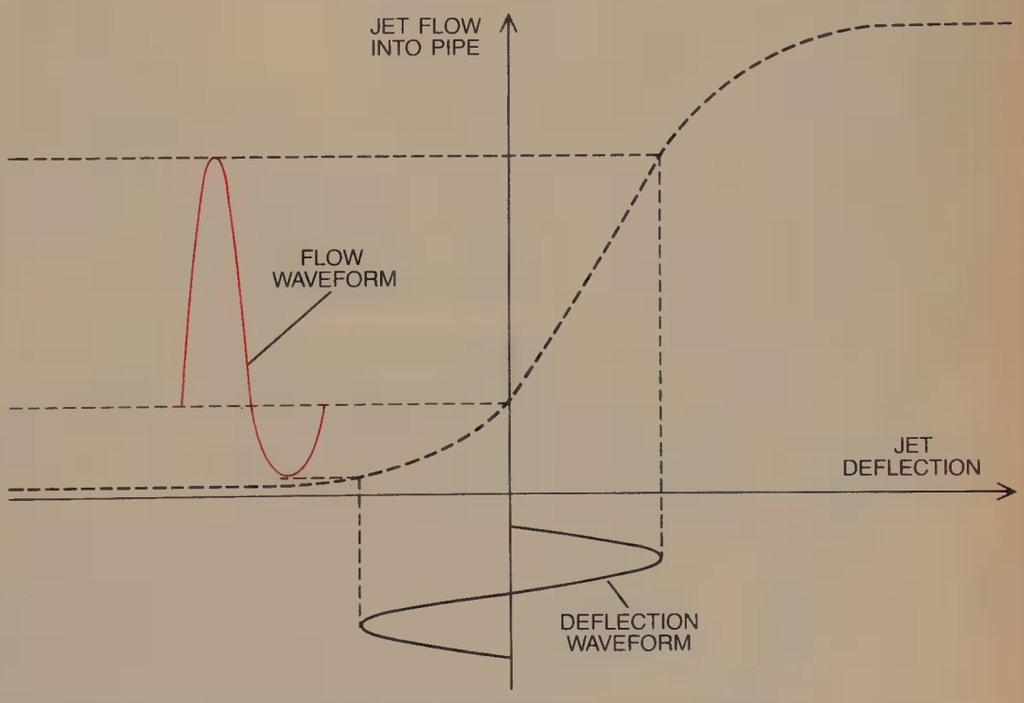
The admittance spiral we have been discussing has an additional complication if the deflection of the jet at the upper lip becomes greater than the width of the jet. The jet then blows nearly completely into and out of the pipe mouth in each cycle and the impulse it gives to the reflected wave in the pipe becomes independent of further increases in amplitude. The efficiency of the jet as a generator decreases accordingly The admittance spiral simply shrinks in size as the amplitude of the jet deflection increases.

The loss of jet efficiency with increasing deflection amplitude is accompanied by increasing energy losses in the organ pipe. The pipe oscillation rapidly settles down at an amplitude for which the energy supplied by the jet exactly balances the energy lost by the pipe. Rather surprisingly it turns out that in most instances the energy losses through turbulence and viscosity are much larger than the losses through sound radiation from the mouth and the open end of the pipe.

The sound of a real organ pipe is not, of course, limited to a single frequency; it has many components of higher frequency. One can demonstrate that these components are all exact harmonics of the fundamental frequency: integer multiples of that frequency. Under steady blowing conditions the shape of the sound wave seen on an oscilloscope remains exactly the same. Even the smallest deviation from exact integer multiples for the frequencies of the components would lead to a slow



ACOUSTIC ADMITTANCE OF THE JET as a function of frequency or of blowing pressure takes the form of a spiral in which the distance from the origin represents the magnitude of the admittance and the angular position represents the phase relation between the acoustic flow out of the mouth of the pipe and the pressure just inside the mouth. When the outflow is in phase with the pressure, the admittance lies in the right half of the spiral and the energy of the jet is being dissipated. For the jet to act as an acoustic generator the admittance must lie in the left half of the spiral, which requires that the back-and-forth displacement of the jet be offset, or delayed, in phase with respect to the pressure inside the mouth of the pipe. The wave reflected from the jet is then larger than the incident wave. The reference angle from which the phase delay is measured depends on which of two excitation mechanisms discussed in the text is dominant, the one proposed in the last century by Hermann von Helmholtz or the one proposed by Lord Rayleigh. When the admittance falls in the upper half of the spiral, the jet lowers the natural resonance frequency of the pipe; when it falls in the lower half of the spiral, the jet raises the resonance frequency. The outer parts of the spiral have been reduced in size for clarity.

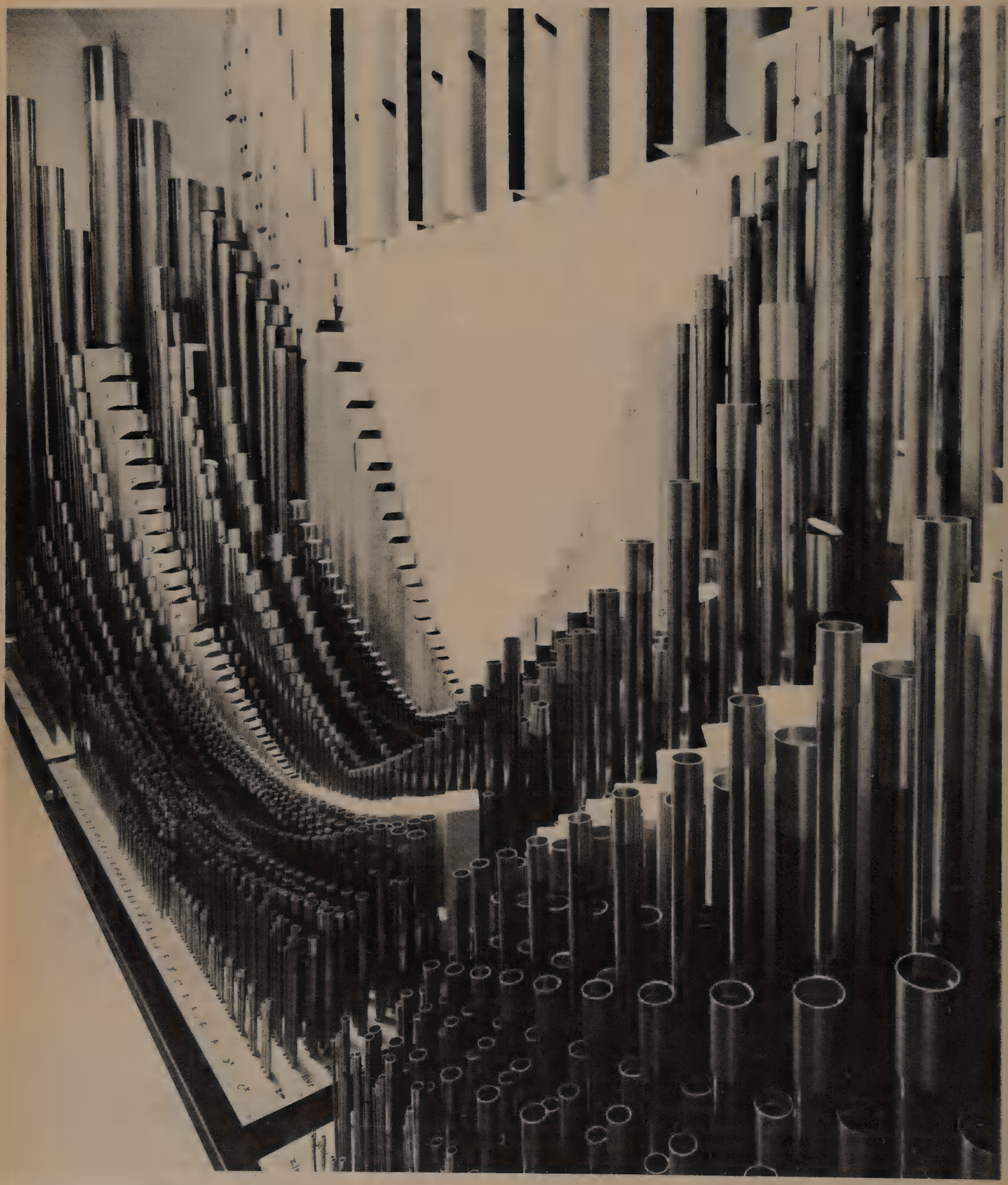


CURVE GIVING THE JET FLOW INTO A PIPE (*broken curve*) for a given deflection is asymmetric around zero deflection because the lip of the pipe has been adjusted to cut the jet slightly away from its central plane. When the jet is deflected in a simple sinusoidal way with a large amplitude (*solid black curve*), the jet flow into the pipe (*colored curve*) "saturates" first at one end of its range, where the jet is blowing completely outside the pipe lip. For a still larger amplitude the jet flow will also saturate at the other end of its range, where the jet is blowing completely into the pipe. The offset of the lip gives the flow waveform an asymmetric shape that has frequency components at all integer multiples of the deflection frequency. It is these harmonic components that are responsible for the rich sound of a well-adjusted organ pipe.

but clearly discernible change in the waveform.

This behavior is interesting because the resonances of the air column in an organ pipe, or indeed in any open pipe, have frequencies that differ slightly from exact harmonic multiples. The reason is that the effective length of the pipe decreases slightly with increasing frequency because of changing acoustic flow at the open ends. As we shall see, the harmonics in the sound of an organ pipe are generated by the interactions of the jet and the pipe lip, so that the pipe acts largely as a passive resonator as far as the upper harmonics are concerned.

The resonances of a pipe develop when there is maximum movement of the air at the open ends of the pipe, which is the same as saying that the admittance of an organ pipe should be at a



INTERIOR VIEW OF ORGAN at the Sydney Opera House shows some of its 26 ranks of pipes, most of which are of metal but some of which are of wood. The length of the speaking part of each pipe doubles at every 12th pipe; the pipe diameter doubles at about every 16th pipe. Through long experience master organ builders arrived at the proportions necessary for achieving balanced tone quality.

maximum at the mouth. This leads to the conclusion that the resonances of a pipe open at its far end occur at frequencies for which the pipe length is an integral number of half wavelengths for sound in air. If the fundamental frequency is f_1, the upper resonances are $2f_1$, $3f_1$ and so on. (Actually all the upper resonance frequencies are stretched a little beyond these ideal values, as we noted above.)

For a pipe that is stopped, or closed, at its far end, resonances occur when the pipe length is an odd number of quarter wavelengths. Therefore a stopped pipe need be only half as long as an open pipe to produce the same note; its resonances are at f_1, $3f_1$, $5f_1$ and so on.

Returning to the organ-pipe jet, we have seen that high-frequency disturbances tend to die out along the jet as the width of the jet increases, so that, almost independently of the higher-frequency components in the acoustic field at the pipe mouth, the tip of the jet at the upper lip moves back and forth almost sinusoidally at the frequency of the fundamental component of the pipe sound. A sinusoidal flow of the jet, however, does not lead to a sinusoidal jet flow into the pipe, because the flow "saturates" by flowing completely inside or outside the upper lip at each end of the jet's deflection range. More than that, the lip is usually offset so that it does not cut the jet exactly along its central plane, thus making the saturation asymmetric. The waveform of the jet flow into the pipe therefore has all the harmonics of the fundamental frequency, locked into precise frequency and phase relation, and the relative amplitudes of the high-frequency harmonics increase rapidly as the amplitude of the jet deflection increases.

In a typical organ pipe the distance the jet is deflected in the mouth is on the same order as the width of the jet at the upper lip. As a result there is a wide spectrum of harmonics in the jet flow. If the jet were to strike the lip exactly symmetrically, the even-numbered harmonics would be absent from the excitation. The jet is usually given a small offset so that all the harmonics are present.

As one might expect, the quality of sound delivered by an open pipe is rather different from that delivered by a stopped pipe. The frequencies of the harmonics in the driving force produced by the jet are exact integer multiples of the fundamental frequency of the wave in the jet. The air column in the pipe will respond strongly to a particular harmonic only if the pipe's acoustic admittance is large, which corresponds to a sharp resonance peak close to the frequency of the harmonic. A stopped pipe, which has only odd-numbered resonance peaks, therefore suppresses all the even-numbered harmonics of its fundamental. The result is a characteristic "hollow" sound in which the even-numbered harmonics are weak, although they are not totally absent. An open pipe, on the other hand, achieves a "brighter" sound because it responds to all the harmonics of its fundamental frequency.

MOUTH OF AN ORGAN PIPE OF THE DIAPASON FAMILY reveals how the languid of the pipe has been "nicked" in order to achieve homogeneous turbulence in the air jet. The pipe is made out of "spotted metal," a tin-rich alloy of lead and tin that solidifies to the surface pattern seen here when it is cast in a metal sheet. The sheet is then formed into a pipe.

The resonance characteristics of a pipe are determined largely by its energy losses. These losses are of two types: (1) viscous and heat-conduction losses to the pipe walls, and (2) radiation losses from the mouth and the open end. The viscous and heat-conduction losses are more important for narrow pipes than they are for wide pipes and turn out to be more important at low frequencies than they are at high frequencies. The reverse is true for radiation losses.

The upshot is that, for a given pipe length and hence for a given fundamental frequency, wide pipes are efficient and well-tuned resonators for only the fundamental and its first few harmonics, which results in a dull, "fluty" sound. Narrow pipes are good resonators for a large range of harmonics, and because higher frequencies are radiated more efficiently than lower ones the tone is thin and "stringy." Between these two extremes are pipes with the bright, ringing sound characteristic of a good organ. These are the pipes called principals or diapasons.

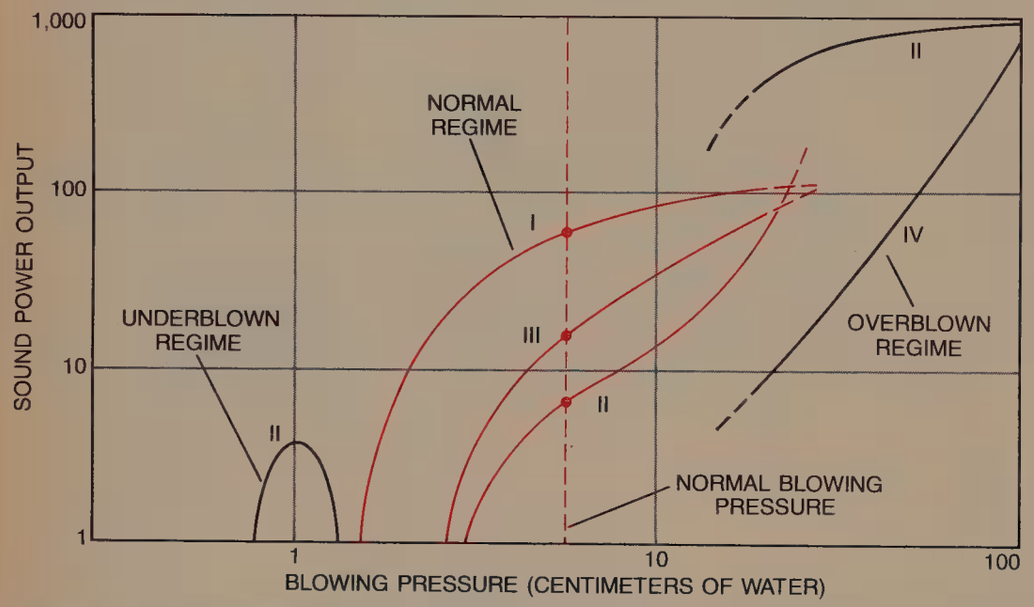
A large organ may in addition have ranks of pipes with a conical body, a perforated stopper or other geometric variations. These features are designed to modify the resonance frequencies of the pipe and in some instances to enhance a narrow range of the higher harmonics to produce a particular tone color. The material from which the pipe is made is of only minor importance.

The fact that the air in a pipe can vibrate in any one of an entire set of possible modes introduces additional complications in the acoustic behavior of pipes. For example, if the blowing pressure in an open pipe is increased so that there is just a quarter of a wavelength of the first mode, f_1, on the jet, the point in the admittance spiral for this first mode will move into the right half and the jet will cease to maintain the mode. At the same time the second mode frequency, $2f_1$, is such that it corresponds to half a wavelength along the jet, so that it can be strongly driven. The sound of the pipe

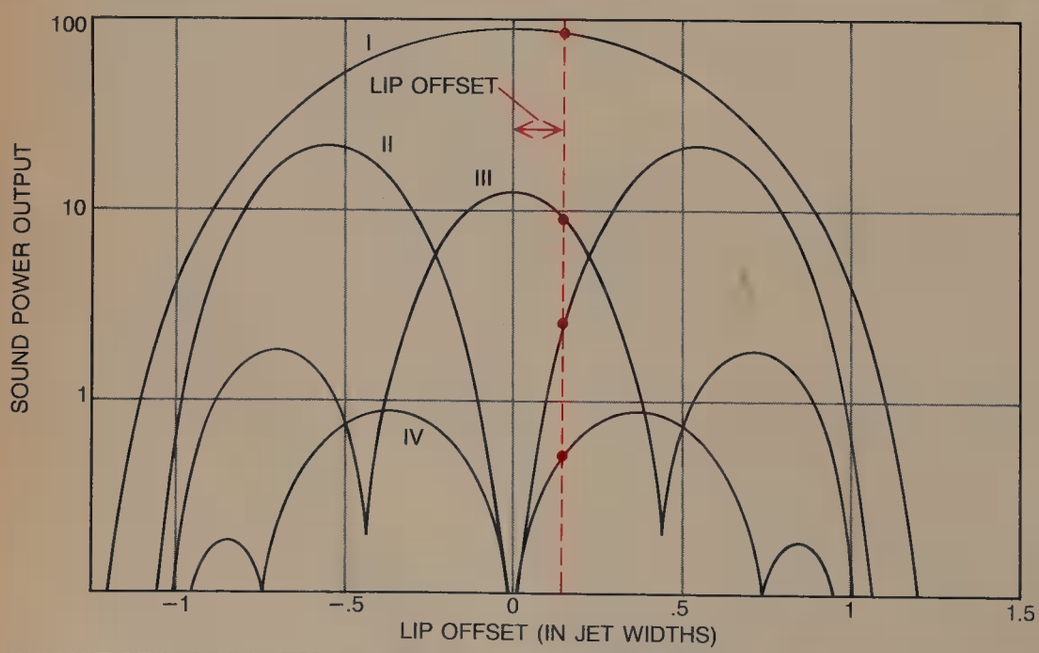
will therefore jump to this second mode, almost an octave above the first, the exact frequency depending on the resonance frequency of the pipe and on the blowing pressure.

A further increase in the blowing pressure can excite the next mode, at $3f_1$, if the cut-up of the pipe lip is not too great. Conversely, it often happens that a low blowing pressure inadequate to drive the fundamental will softly drive one of the higher modes in the second loop of the admittance spiral. These overblown or underblown pipe sounds, although they are interesting in the laboratory, are not exploited in pipe organs except for special effects.

The main problem facing the organ builder who has created a single pipe of the appropriate sound quality is to design a rank of pipes covering the musical compass of the keyboard and matching one another in loudness and harmonic content. A set of geometrically similar pipes, simply scaled up or down in size, does not meet this requirement: the pipes differ in how wall and radiation losses affect different frequencies. In order to achieve constant behavior across the acoustic spectrum one must apply a scaling rule. The diameter of the pipe is varied with the length of the pipe raised to a suitable power, k, whose value is less than 1. The result is that the long bass pipes are made somewhat narrow. The theoretically determined value of k is 5/6, or .83, but when the psychophysics of human hearing is taken into account, a value of .75 proves to be more satisfactory. That value turns out to be close to the empirical rule developed by the great organ builders of the 17th and 18th centuries.



EFFECT OF CHANGING BLOWING PRESSURE on sound output is depicted for a typical organ pipe. The Roman numerals represent the first few harmonics. The central regime (*color*) covers the range of normal speech, with a good balance at normal blowing pressure. At higher pressures the pipe overblows to its second mode; at low pressures there is a weak second mode.

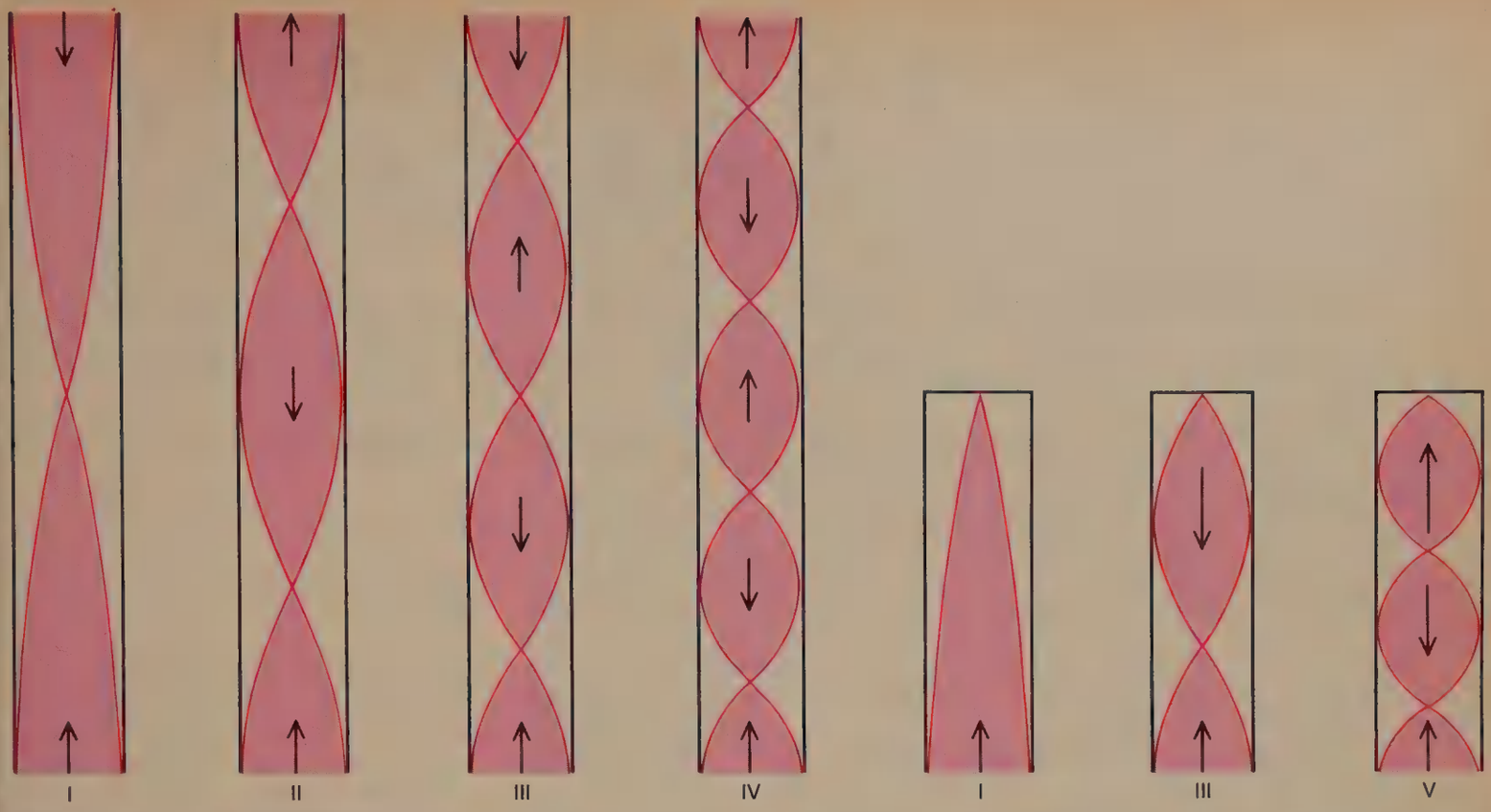


EFFECT OF LIP LOCATION on harmonic components shows the desirability of an offset. If the lip were exactly centered on the jet, the pipe would deliver only the fundamental frequency (*I*) and the third harmonic (*III*). With the lip offset at the location shown by the broken line, the second and fourth harmonics are introduced, greatly enriching the quality of the sound.

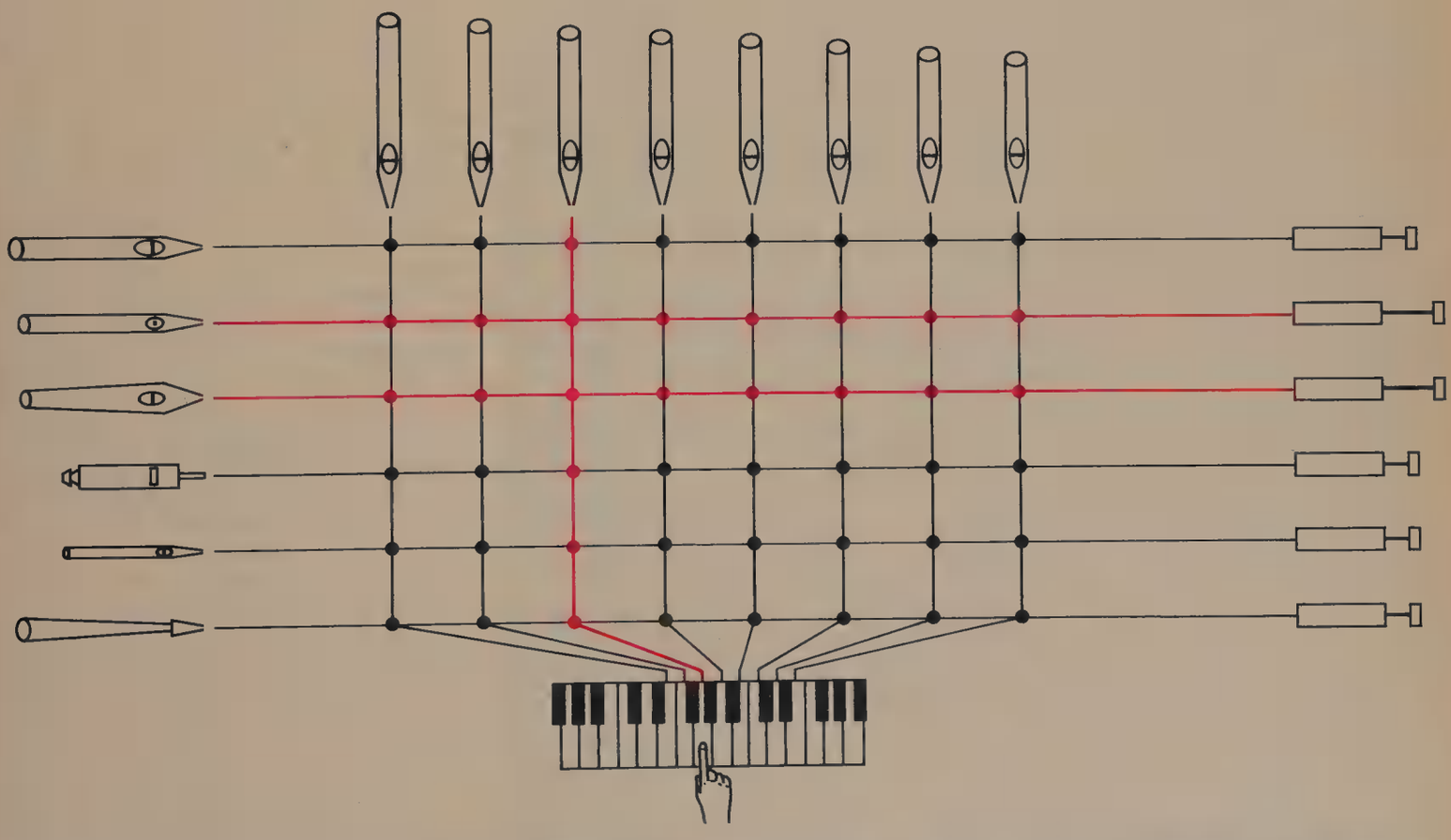
Finally, and of course important from the viewpoint of the player, comes the mechanism by which the speech of the multitude of pipes in a large organ is controlled. The basic design is delightfully simple and resembles the rows and columns of a matrix. The pipes are laid out in ranks, corresponding to the rows of the matrix. All the pipes in each rank have one tone quality, with one pipe for each note of the keyboard or pedal board. The air supply to each rank is controlled by a drawstop knob with the name of the rank written on it. The air supply to the pipes associated with each note, or column of the matrix, is controlled by the keyboard key for that note. A given pipe will sound if and only if it is in a rank for which the drawstop is pulled and on a key channel for which the key is depressed.

In these days one can imagine a host of ways of implementing such a scheme based on digital logic and electrically operated valves under each pipe. Early organs had simple mechanical levers and pallet valves to let air into the key channels, and perforated mechanical sliders to control the entry of air into an entire rank of pipes. Apart from having the virtues of simplicity and durability, this mechanical system gave the player precise control over the rate of opening of the key-channel valve and a feeling of closeness to the operation of an otherwise rather mechanical instrument.

In the 19th century and the early 20th large organs were built with all manner of electromechanical and electropneumatic mechanisms, but recently the emphasis has shifted back to mechanical action for the keys and the pedals, with sophisticated electronic controls for setting stop combinations as the instrument is being played. The organ in the concert hall of the Sydney Opera House, for example, is the largest organ in the world to have a mechanical action. Completed in 1979, it has 10,500 pipes in 205 ranks distributed across five keyboards and a pedal board. The key action is mechanical but is duplicated by an electrical action so that electric couplers can be used. Thus a live performance can be recorded digitally and the tape can be subsequently used to operate the organ, thereby re-creating the original performance. The stops and combination pistons are all electric or electropneumatic and are under the control of microprocessors having versatile stored-program capabilities. In such ways the best resources of modern technology, combined with traditional design and the skills of individual craftsmen following principles known for centuries, create the flexible and glorious sound of a large organ.



STANDING-WAVE RESONANCE PATTERNS are represented schematically for pipes open at both or closed at their upper end. The width of each colored pattern corresponds to the acoustic vibration amplitude in various parts of the pipe. The actual air motion is parallel to the axis of the pipe. Arrows show the direction of air motion during one half of the vibration cycle; during the other half the direction is reversed. The Roman numerals are the harmonic numbers and are proportional to the vibration frequencies. An open pipe has resonances for all the harmonics of its fundamental. A stopped pipe need be only half as long to produce the same note but has resonances for only the odd harmonics. Vibration patterns actually extend slightly beyond the ends of the pipe (calling for the end correction), and the complex geometry of the pipe mouth distorts the patterns somewhat at the lower end of the pipe without altering their "basic" nature.



PIPES OF AN ORGAN are arranged like the rows and columns of a matrix. In this simplified diagram each row, which is called a rank, consists of pipes of the same type with one pipe for each note (*top*). Each column, which is associated with one note on the keyboard (*bottom*), provides access to all the pipes of different kinds (*left*). The drawstops of the organ console (*right*) admit air to all the pipes of a rank; the keys of the keyboards admit air to all the pipes in the key channel for that note. The mechanism is arranged so that air reaches a pipe only if both its row and its column are activated. In the situation that is depicted in the diagram just two pipes will sound.

The Origin of the Cosmic X-Ray Background

What accounts for the diffuse glow of X-radiation that appears to fill the universe uniformly in all directions? New findings suggest that one source may be a multitude of distant quasars

by Bruce Margon

A fundamental observation about the structure of the universe can be made with the unaided eye on any clear night: The space between the stars is black. The fact that most of the visible light in the night sky emanates from individual stars, and that these light sources are in turn grouped into galaxies, is one of the most basic findings in cosmology. An observer surveying the heavens at wavelengths in the ultraviolet, infrared or radio regions of the electromagnetic spectrum finds this same clumping of the constituents of the universe. At the much shorter wavelengths of X-radiation, however, the situation is quite different. Since 1962 it has been known that every part of the sky emits a uniform glow of X rays. After two decades of intensive study the origin of this diffuse X-ray background is still a subject of controversy.

The sky is uniformly bright at only one other range of wavelengths: in the microwave region of the spectrum. It is a curious accident of astronomical history that the microwave background radiation was discovered in 1965, soon after the first observation of the X-ray background. The origin of the microwave background has in the meantime been identified, and the phenomenon is widely recognized as a profound clue to the evolution of the universe. Most astrophysicists regard the microwave glow as a remnant of the big bang, the massive explosion that began the current expansion phase of the universe.

X rays and microwaves differ in wavelength by a factor of about a million, and in nature the two kinds of electromagnetic radiation are normally generated under very different physical conditions. Accordingly there is no reason to expect any direct connection between the two kinds of cosmic background radiation. Nevertheless, the microwave background stands as both a tantalizing hint and a nagging frustration to students of the X-ray background. The hint is that a better understanding of the origin of the X-ray background, although still elusive, may eventually reveal clues to the large-scale structure of the universe that are just as profound as the inferences drawn from the microwave background. The frustration is that although the study of both phenomena began at roughly the same time, the microwave background is now understood in some detail, whereas there is no general agreement on even the basic physical mechanism underlying the X-ray background.

This uneven progress in deciphering the two kinds of cosmic background radiation can be attributed in part to the different observational techniques employed at the two wavelengths. Microwaves, which range in length from about .1 centimeter to 10 centimeters, can readily penetrate the earth's atmosphere and hence can be detected by ground-based antennas. In the spectral region where the X-ray background was discovered, between one angstrom and 10 angstroms (that is, between a 100-millionth and a 10-millionth of a centimeter), the earth's atmosphere is totally absorbent. Astronomical observations at these wavelengths therefore require that the detector be placed completely above the atmosphere. German V-2 rockets captured by the U.S. at the end of World War II gave American astronomers their first opportunity to lift astronomical instruments into space. A rather primitive X-ray detector carried on one of those early flights revealed that the solar corona, the extremely hot, dilute outermost layer of the sun, is an identifiable source of X-radiation.

In one sense the discovery of X rays from the sun was a triumph. It marked the advent of a new observational technique in astronomy. In another sense it was a disappointment. The intensity of the solar X rays is extremely feeble: about a millionth of the sun's visible-light output. Armed with this ratio, and with the knowledge that the sun is a normal star, one can easily compute the expected X-ray emission from the roughly 100 billion other stars in our galaxy. The calculation shows that even the nearest of the sun's neighboring stars would be expected to deliver 40 billion times less X-radiation to the earth than the sun does. In other words, X rays are detected from the sun not because it is a remarkable object but only because it is a nearby one. Thus the first achievement of X-ray astronomy—the detection of X rays from the sun—seemed to indicate that X-ray emission would be a property of very limited usefulness in stellar astronomy, at least for the study of the vast majority of normal stars.

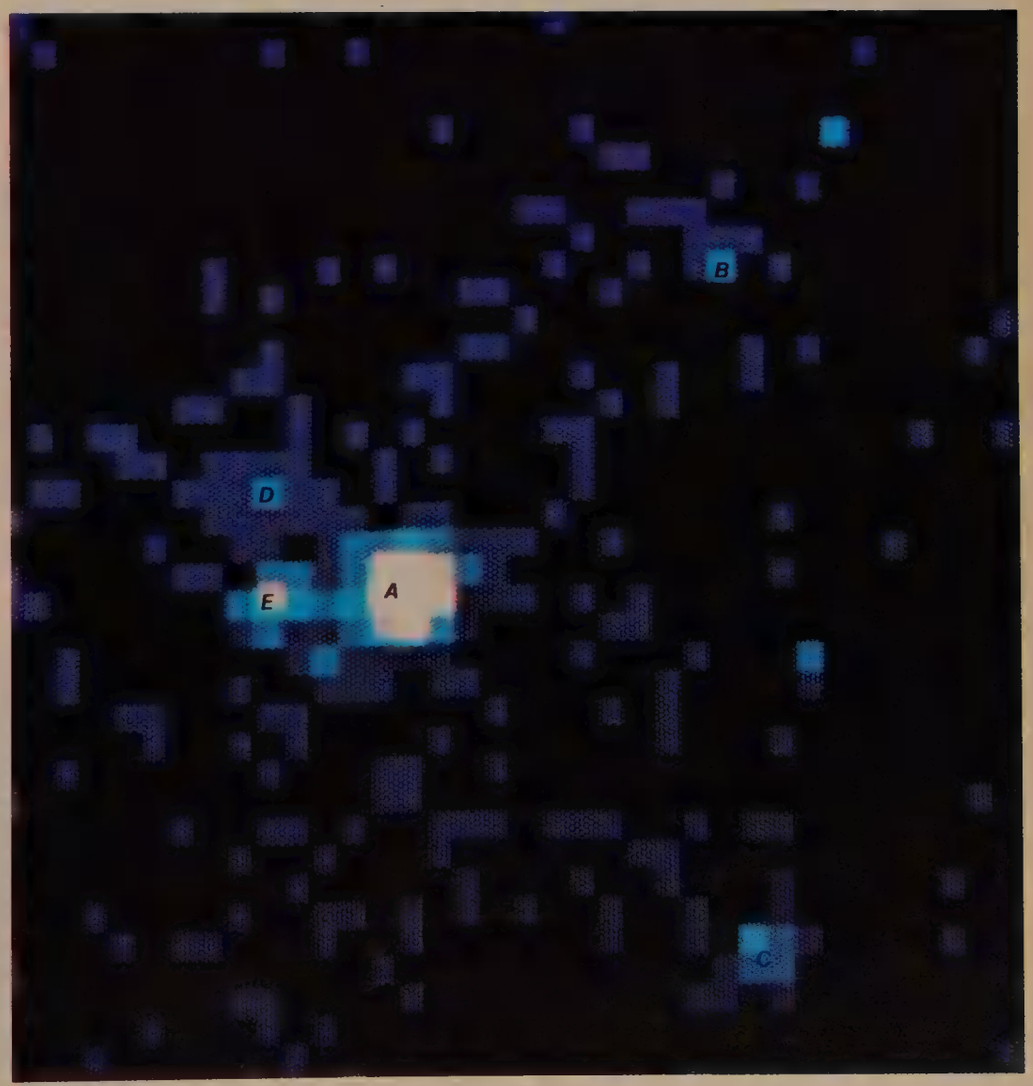
Fortunately this argument did not deter Riccardo Giacconi and his colleagues at American Science & Engineering, Inc., from exploring the potential role of X-ray observations in the study of abnormal celestial objects. In 1962 they built an improved X-ray detector that was briefly lifted above the atmosphere on board a small rocket. The new detector was substantially more sensitive than the X-ray detectors that had been carried aloft to observe the sun; it was the first such device flown that had sufficient sensitivity to detect X-ray sources outside the solar system, if indeed any such sources existed.

Giacconi's experiment resulted in two spectacular discoveries. First, a discrete point source of X rays was found in the direction of the constellation Scorpius. Although the X-radiation from this source was a million times fainter than that from the sun, it was still much more intense than the predictions had indicated the X-radiation from even the nearest normal star could be. The newly discovered object, designated Sco X-1, was only approximately located in the sky by the initial observation, and so it was not possible to determine whether the source of the X-ray emission corre-

sponded to a visible object. Successive generations of experiments flown by groups of astronomers from a number of countries have since found several hundred such X-ray-emitting objects in the galaxy and have analyzed them in increasing detail, to the point where these galactic X-ray sources are now understood quite well.

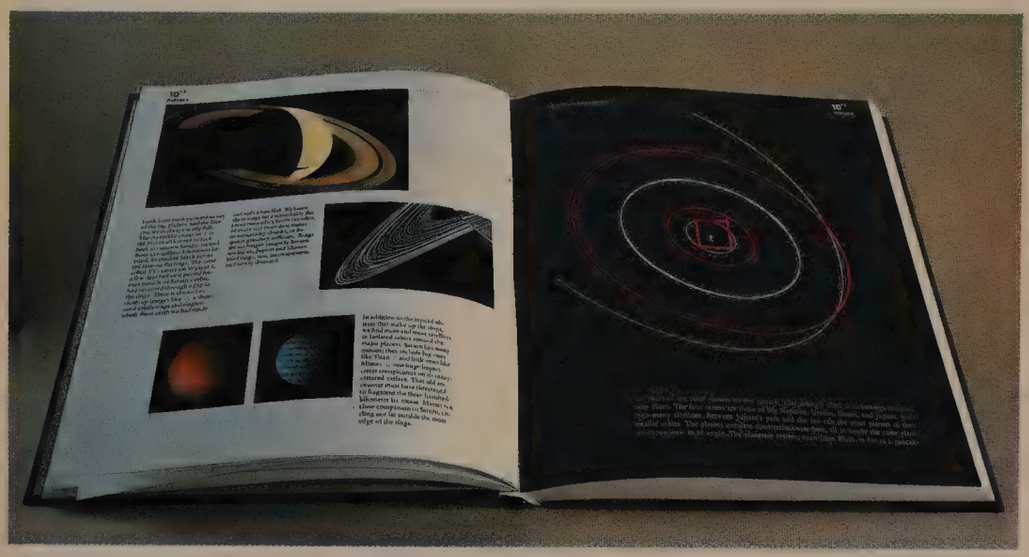
The prediction that it would not be possible to detect X rays from more than a few of the nearest normal stars turned out to be correct. X-ray sources such as Sco X-1 are much more distant and distinctly abnormal stellar systems. The X-radiation from these sources is generated by a process that involves the exchange of stellar material between a fairly ordinary star and a highly evolved, compact star orbiting nearby; the compact companion in such an X-ray-emitting binary system may be a white dwarf, a neutron star or possibly even a black hole.

The few minutes of data from this first "extrasolar" X-ray experiment provided a surprise in addition to the dis-



QUASARS AND NORMAL STARS appear quite similar in an ordinary, visible-light photograph (*top*), but they can be readily distinguished with the aid of a corresponding X-ray image (*bottom*). The photograph, reproduced here in negative form to help preserve fine details, was obtained by means of the 1.2-meter Schmidt telescope on Palomar Mountain. It shows a region of the sky measuring about half a degree on a side (approximately the angular diameter of the full moon) in the constellation Hercules. The photograph includes about 1,000 starlike images, the vast majority of which are normal stars in our galaxy. Accordingly any selection technique designed to identify the small number of quasars likely to be found in any such region must reject the foreground stars with very high efficiency. The false-color X-ray image of the same region was made from data recorded by an X-ray detector on board the orbiting satellite known as the Einstein Observatory. Only five X-ray sources are clearly discernible, four of which can be associated with quasars. The source labeled *A* is a previously known quasar that served as the target of this particular X-ray exposure; designated 3C 345, it is estimated to be some 7.2 billion light-years from the earth. The other four X-ray sources were discovered on this image. Based on spectroscopic observations made with large telescopes, the author and his colleagues were able to determine that the serendipitously discovered X-ray sources labeled *B, C* and *D* are in fact previously uncatalogued quasars. Sources *B* and *C* are respectively 9.5 and 6.7 billion light-years away; source *D* is the combined emission from two separate quasars, one at a distance of 7.9 billion light-years and the other at 7.2 billion light-years. Source *E* coincides with the comparatively nearby but abnormal galaxy NGC 6212; it is only about 560 million light-years away. The optical counterparts of the X-ray sources are indicated on the photograph. The X-ray image was provided by Gary A. Chanan of Columbia University.

SCIENTIFIC AMERICAN LIBRARY



MAN'S CORNER OF THE UNIVERSE lies open here at the pages in *Powers of Ten* that display the universe at the 10^{13} meters dimension (10 billion kilometers), about the diameter of the solar system.

POWERS OF TEN

A wondrous journey across the universe from its largest dimension to its smallest

Here is an atlas of the universe and a handbook to its exploration by the disciplines of science. In vivid color illustration—from the informed imagination and from the instruments of science—*Powers of Ten* shows how the universe looks from the largest to the smallest dimensions now in the grasp of human understanding.

This spectacular journey, with its sweeping changes in perspective, proceeds in ordered steps, each a multiple of ten of the last step or of the next.

There are 42 of these equal logarithmic steps from the outermost regions to the innermost recesses of the known universe. More than half were taken in the lifetime of scientists now alive.

The visual magic of this first-time-in-history journey is the bold invention of Charles and Ray Eames, who have here turned their mastery of multi-media technology to book design. Your tour guides are Philip and Phylis Morrison. They supply instructive and poetic pages of commentary, answering questions and showing how the sciences get those answers at each of the 42 steps.

The SCIENTIFIC AMERICAN LIBRARY publishes *Powers of Ten* as its premier volume because it sets out the agenda of the books that are to come. Each succeeding book further illuminates the ideas and images displayed, step by step, on the pages of *Powers of Ten*.

The SCIENTIFIC AMERICAN Library opens an entirely new approach to understanding and sharing in the work of science. Each specially commissioned volume

. . . comprehends a large topic that engages more than one discipline and, typically, recurs in many articles in SCIENTIFIC AMERICAN;

. . . is written by a scientist who has made his own pivotal contribution to the work on the topic;

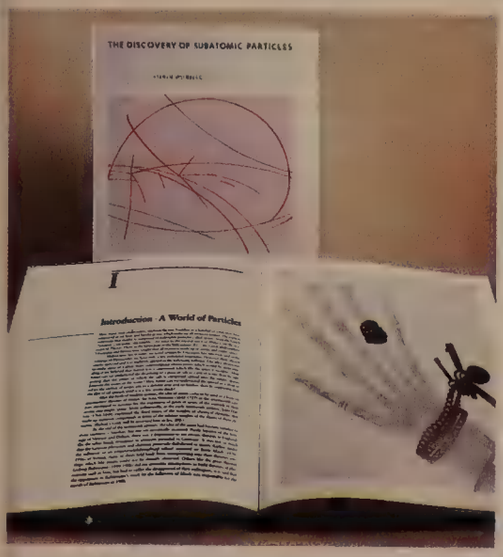
. . . rewards its reader not with "information" but, thanks to the sympathetic collaboration of the author and editor in the preparation of text and illustration, with command of the concepts central to understanding of the work and appreciation of its meaning.

Readers of this magazine are invited to accept membership in the SCIENTIFIC AMERICAN Library. Members are entitled to receive for inspection, without any charge, each volume as it is published. Membership, cancellable at any time, will gather in time a reference library of the member's own choosing, with a reach and accessibility unprecedented in the literature of science.

A SAMPLING OF THE DIMENSIONS of the known universe explored on the journey through the Powers of Ten is shown here. At center, the Taj Mahal adorns the 10 meter world in reach of our unaided senses. Dimensions grow smaller to the left: first, crystalline skeletons of single-celled diatoms at 10^{-4} meter (.1 millimeter) and at far left, tungsten atoms, 10^{-10} meter in diameter, brought into visibility by a field-emission microscope. At right, looking across the Indian Ocean toward the South Pole through the window of Gemini II, the curvature of the Earth, 10^{7} meters in diameter. At far right, a quasar, 10^{25} meters away (one billion light-years), caught by a radio telescope and imaged by computer graphics.

SUBATOMIC PARTICLES

by Steven Weinberg



DISSECTION OF ATOM began with accidental discovery of X-rays by Roentgen in 1895.

As everyone knows, the once irreducible atom has been shown to be composed of still more fundamental particles: the electron, proton and neutron. But how do we know this? In *The Discovery of Subatomic Particles* the reader comes to know the answer just as Steven Weinberg and other physicists did, thereby sharing in the intellectual enterprise that has reshaped physics in the twentieth century.

We do not accept scientific truth on faith or authority. We know truths of this kind are grounded on experiment and observation—operations we might have performed ourselves, but in any case can perform inside our heads. With Weinberg the reader of this book reenacts the historic succession of experiments that disassembled the atom. From each experiment to the next, the reader will gain a confident understanding of the laws of physics that govern the events encountered.

HUMAN DIVERSITY

by Richard C. Lewontin

The relevance of science to human values finds compelling demonstrations in the proper understanding of the genetics of our species. In *Human Diversity* Richard Lewontin shows that each human being differs from all others because of the interaction of genetic differences, environmental differences, and chance events that occur during development. There is no unique "normal" genetic constitution.

The genetic markers that supposedly divide the species into races are only a trivial fraction of the total spectrum of diversity in which each person finds singularity. The genetic endowment held in common across racial, national and cultural boundaries is a consequence of the biological history of our species. The segregation of people into social castes can never again invoke genetics as its rationalization.

FOSSILS

by George Gaylord Simpson

One of the architects of the modern evolutionary synthesis here establishes the nature—and the true splendor—of the solid evidence upon which much of the theory and the facts of evolution rest. Simpson declares: "The primary record of the history of life is written in the successive strata of rocks as in the pages of a book. Fossils may be called the writing on those pages. Traces of organisms living at successive geological times, they represent once living things and should be seen as such. They must be put in their sequence in time. They were influenced by and bear witness to geographic and geological changes on the earth. They are basic materials for the study of organic evolution. Their study combines historical geology and historical biology into one great synthesis."

THE SOLAR SYSTEM

by Roman Smoluchowski

The fine-grained picture of the Earth and its immediate cosmic neighborhood now in the possession of mankind is here comprehensively assembled. The author has brought together the latest graphic images from the satellites and from earthbound telescopes as well. (We know the terrain of cloud-wrapped Venus from radar echoes returned to the 600-foot radio telescope at Arecibo in Puerto Rico.) He shows us our solar system in time as well as in space. In the wealth of images of the Sun, of the planets, of their satellites, of the swarming asteroids and of the cometary messengers from the primordial ice clouds on the outermost circumference of the system, we comprehend the origin, the history and the fated future of our Sun and its planets.

MUSICAL SOUND

by John R. Pierce

In this volume, the author will share with you his joy in the discovery that understanding from physics and mathematics enlarges and enhances one's appreciation of music. Thereupon you will find yourself in the company of a long succession of natural philosophers who have pursued just that strategy.

John R. Pierce is the latest of these natural philosophers. To his work in music he brings his study of sound, its transmission and reproduction, at Bell Telephone Laboratories plus the insights of such contemporary composers as Pierre Boulez, John Cage and Edgard Varèse.

In the live recordings that come with this volume Pierce demonstrates something of what the psychology of acoustics has learned about the perception, the illusion and the effect of sound. He imbues the reader with his confidence that rational enquiry into this most intensely subjective of all aesthetic experiences will open up new realms of sound to that experience.

SOUND SHEETS, for 33-r.p.m. turntable, demonstrate acoustic illusions discussed in book.

Charter Membership Reservation Form

To SCIENTIFIC AMERICAN LIBRARY
Holmes, Pennsylvania 19043

Please enroll me as a Charter Member of the SCIENTIFIC AMERICAN LIBRARY. Send me the Premier Volume — *Powers of Ten* — by return mail. Thereafter, I am entitled to receive for inspection a new volume every two months, as it is published, for fifteen day's *free examination.*

I need send no payment now. For each volume I keep I will be billed $19.95 plus postage and handling. This price is guaranteed for at least one year, regardless of any price increases asked of new members.

I understand that I may cancel my Charter Membership at any time for any reason.

NAME _____
(Please print clearly)

ADDRESS _____

CITY _____ STATE _____ ZIP _____

Canadian residents should pay in U.S. dollars.

JD

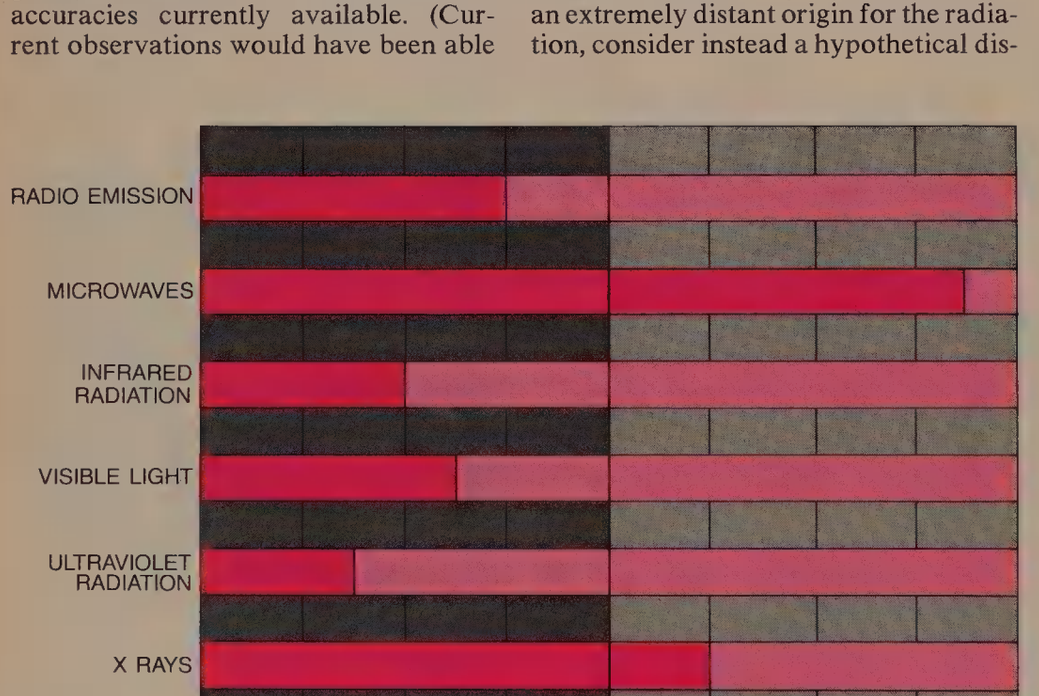
covery of Sco X-1. The rocket carrying the X-ray detector spun quickly during its brief ascent and descent, thereby sweeping out a substantial fraction of the celestial sphere. Contrary to expectations, the detector registered an apparently constant intensity of X-radiation at all times during the observations above the atmosphere, regardless of the viewing direction of the rocket. Simultaneous with the beginning of stellar X-ray astronomy, therefore, was the discovery of the diffuse X-ray background.

Like the stellar X-ray sources, the X-ray background has been the object of study by almost all subsequent experiments in X-ray astronomy. The first of these efforts called for additional rocket flights; for the past 10 years or so the detectors have been carried on board orbiting satellites. Several fascinating features of the X-ray background have emerged from these studies. Perhaps the most remarkable is the uniform intensity of the radiation from different parts of the sky. Every region of the sky exhibits an X-ray brightness that is precisely equal to that of neighboring regions, to within the best measurement accuracies currently available. (Current observations would have been able to detect fluctuations in brightness of about 1 percent between adjacent regions of the sky with an angular extent about the size of the full moon.) In other words, the X-ray background is exceedingly isotropic: whatever generates the background radiation produces an equal number of X rays from every direction in the sky, to an accuracy exceeding 99 parts in 100. (The microwave background radiation is also known to be highly isotropic, to an even greater measurement accuracy: about 2,999 parts in 3,000.)

The near-perfect isotropy of the X-ray background is clearly a vital clue to its origin. The clue is usually interpreted as indicating that the source (or sources) of the radiation is either very near or very far but not at some intermediate distance. The possibility that the source is very near is easy to understand; an observer completely immersed in the radiation from a nearby source would perceive a perfectly isotropic glow. In the case of the X-ray background, however, it is difficult to postulate any highly uniform source of X rays in the vicinity of the solar system.

To understand why the alternative is an extremely distant origin for the radiation, consider instead a hypothetical distribution of X-ray emitters at some intermediate range of distances, for example at distances comparable to the spacing between the stars in our galaxy or even between nearby galaxies. In that case there would have to be one or more emitters in some direction that were closer to the solar system than all the other emitters. When one looked in that direction, one would expect to see a bright patch corresponding to the nearer source of radiation.

This conclusion, however, would contradict the original observation that the X-ray background is highly isotropic: there are apparently no lumps of X-ray-emitting matter in any part of the sky. The only reasonable explanation is that most of the radiation comes from enormous distances, distances that are in fact comparable to those of the most distant known objects in the universe. As with the microwave background, the eventual explanation of the diffuse X-ray background will almost surely have some bearing on the cosmological question of the large-scale structure and evolution of the universe.

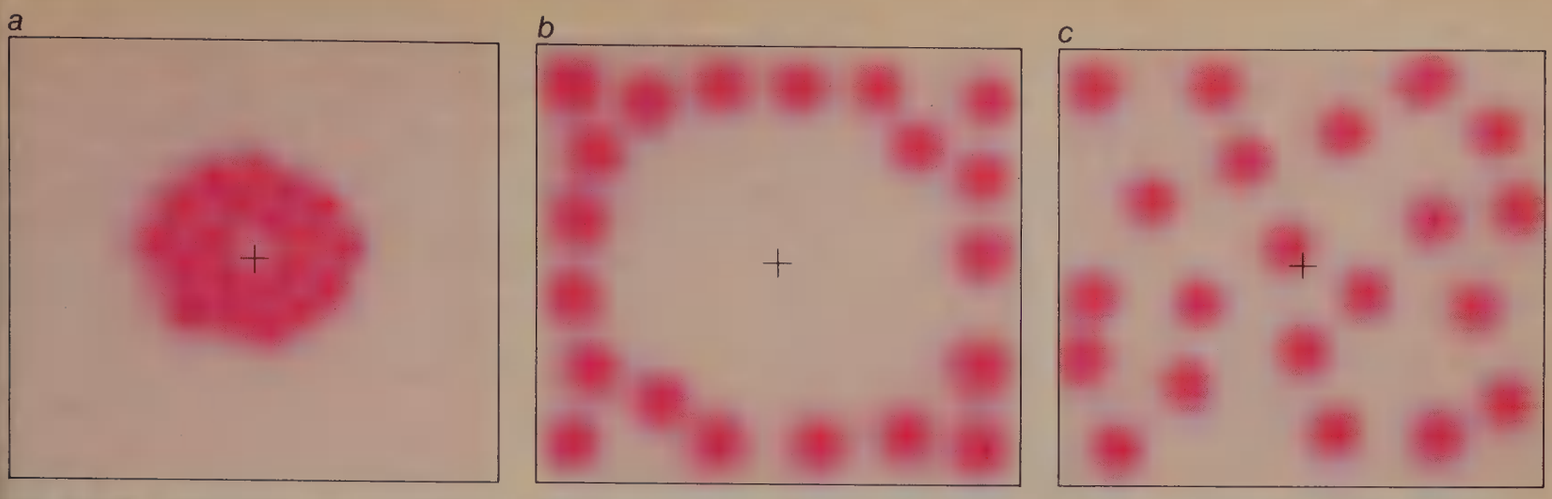


RATIO OF ELECTROMAGNETIC ENERGY measured by a hypothetical observer outside our galaxy to that known to be present within the galaxy is represented by the colored bars for different parts of the spectrum. At five of the seven spectral ranges shown the ratio is less than 1, indicating that the energy per unit of volume at these wavelengths is greater within the galaxy than it is outside. In other words, the outside observer would perceive that the dominant source of the radiation at these wavelengths is concentrated into galaxies, with little radiation being emitted from the space between galaxies. At two wavelength ranges, however, the reverse is true: the energy of the radiation that fills the universe is greater than the energy of the radiation within the galaxy at the same wavelengths. In the latter case the outside observer would have difficulty distinguishing galaxies, since they would be overwhelmed by the diffuse background radiation. Unlike the microwave background radiation, which is generally interpreted as being a remnant of the big bang, the X-ray background is still a mystery. The chart was adapted from one devised by Joseph Silk of the University of California at Berkeley.

A second clue to the origin of the X-ray background arises from the fact that it is X-radiation that is generated, rather than some longer-wavelength radiation. X-radiation is one of the most energetic forms of electromagnetic radiation; hence the source of the observed emission must be some highly energetic process. Furthermore, the process must be one that operates in conditions prevalent on the scale of the universe.

One such process for the generation of X rays has been familiar to physicists for decades; it involves the extremely hot, dilute gas called a plasma. In such a gas the electrons and the protons, which are normally bound together to form atoms, are present as an unbound mixture of rapidly moving particles. Since an electron and a proton have opposite electric charge, there is an attractive force between nearby pairs of electrons and protons in the plasma, and near-misses of free electrons attracted to free protons but not colliding with them are common. Since the advent of quantum physics early in the century it has been known that a free electron passing close to a free proton loses some energy, and that the lost energy is carried away by a discrete packet of electromagnetic energy: a photon. The radiation created by the interaction of an electron and a proton is called bremsstrahlung (German for "braking radiation").

The exact wavelength (or range of wavelengths) of the photons created by the bremsstrahlung process is sensitively related to the temperature of the plasma. For a plasma with a temperature in the range between 10 and 500 million degrees Kelvin bremsstrahlung is expected to generate X rays with wavelengths



EXTREME ISOTROPY, or uniformity, of the X-ray background implies that it originates either very close to the earth or very far from it but not at some intermediate distance. If the sources of the radiation are all nearby (*a*), an observer would be completely immersed in the radiation and would therefore see a uniform glow. If the sources are all extremely far away (*b*), the distance from the observer to various sources may differ slightly, but the difference would be small compared with the enormous average distance and therefore no bright patches would be observed. If the radiation originates at a variety of intermediate distances (*c*), however, the difference between a nearby source and a distant source would become important; one would then expect to see some part of the sky, corresponding to the nearer source, as a bright patch. Various lines of evidence point to an extremely distant origin for the diffuse X-ray background, although it is still uncertain whether the source is actually a diffuse gas, an ensemble of discrete objects or some combination of the two.

predominantly between one angstrom and 10 angstroms. Is there any reason to expect that the universe is more or less uniformly filled with a very hot plasma that generates by the bremsstrahlung process the radiation observed today as the X-ray background? There are a number of reasons to suspect the existence of just such a hot gas.

The space between the stars in the galaxy, although it is an excellent vacuum by terrestrial standards, contains a significant amount of dilute gas. The interstellar medium accounts for only about 10 percent of the mass represented by all the stars in the galaxy, but it plays a crucial role in galactic evolution: it is the raw material out of which new stars form. Astronomers have long sought to understand how, by analogy, the enormous agglomerations of matter that ultimately become individual galaxies form. It seems likely, although it is by no means certain, that galaxies condensed from a dilute gas that filled the entire universe shortly after the big bang: the intergalactic medium.

The hypothetical nature of this scenario must be recognized; although the interstellar medium can be observed directly and star formation is known to be a continuing process in our galaxy, the intergalactic medium has not been observed directly and the formation of galaxies apparently stopped throughout the universe billions of years ago. Nevertheless, there is reason to suspect that some intergalactic medium may still exist. Few processes in nature work with an efficiency of 100 percent, and so it would be surprising if the process of galaxy formation swept away every atom of the original medium.

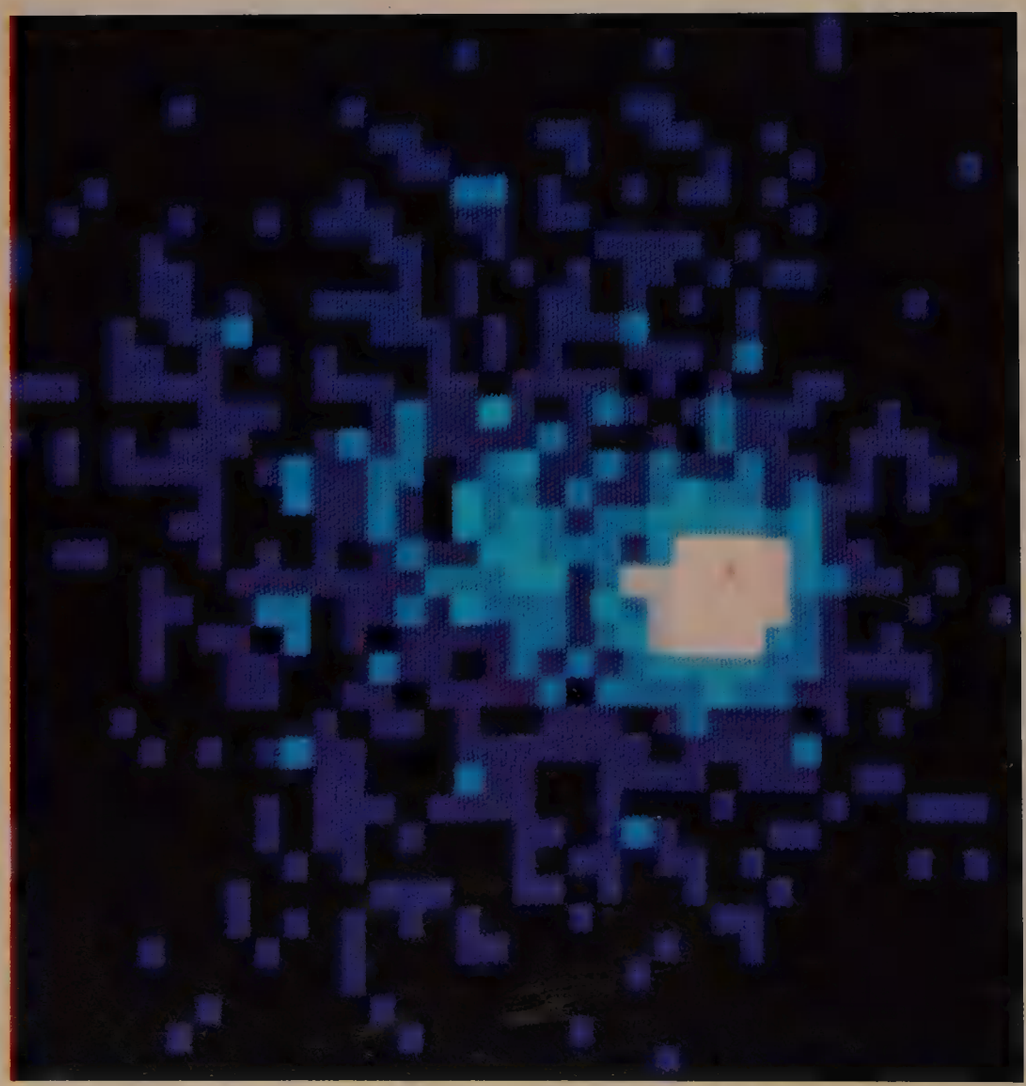
If the diffuse X-ray background is to be explained as bremsstrahlung from an intergalactic medium, the question arises of whether it is reasonable to suppose such a plasma would have the extremely high temperatures required to generate X rays. There are two considerations that indicate this assumption may be plausible. The first is an argument by elimination. If the intergalactic medium exists, and if the gas were as cold as the interstellar medium, then such a gas, intervening between our galaxy and very distant objects, would give rise to noticeable absorption features in the visible-light spectrum of those distant objects. Such absorption features, which can be shown to have a very specific signature, have been thoroughly searched for, and they are normally absent. If there is any cold intergalactic gas, it is certainly not ubiquitous enough to fill the universe isotropically. Hence by default one is led to suspect a hot intergalactic gas; if such a gas exists in any significant quantity, there is strong observational evidence that it is not cold.

The second plausibility argument concerning a hot intergalactic gas involves a calculation of how long the gas could remain hot. If such a medium were expected to cool rapidly compared with current estimates of the age of the universe, one would be faced with the awkward situation of having to explain not only how the gas reached such a prodigious temperature originally but also what mechanism continues to reheat the gas throughout the lifetime of the universe, making the X-ray background observable today. Actually it is easy to calculate that in such a dilute plasma the cooling time would be comparable to or greater than the age of the universe.

Bremsstrahlung is simply not an efficient way of extracting energy from such a thin medium, and there is no other likely cooling mechanism. Assuming there was a way of heating the intergalactic gas once, earlier in the evolution of the universe, one would therefore expect the gas to still be hot today. A number of such possible heat sources are thought to have been present in the violent early stages of the universe. Although the temperature of the gas is very high, it can be shown that the amount of the gas in any one small volume of the universe is likely to be minuscule. Thus the total energy content of the gas, and hence the energy needed to be supplied by this ancient heating mechanism, is large but not overwhelming.

There is even existing observational precedent for significant amounts of very hot cosmic plasma, although not in a form as omnipresent as would be required for an intergalactic medium. Throughout the universe it is commonly found that individual galaxies tend to gather into large, gravitationally bound clusters. The richest clusters can have hundreds of, or even a few thousand, member galaxies. One of the early discoveries of X-ray astronomy was that rich clusters tend to be detectable sources of X rays. Detailed studies of the X rays from clusters have shown that the emission is not due to the sum of the individual sources within the member galaxies. Instead the emission is extended and diffuse, filling the space between the member galaxies and often having a spatial extent comparable to that of the entire cluster. This X-ray emission is now attributed to bremsstrahlung from a hot plasma (typically between 20 and 60 million degrees K.) that fills the volume of the cluster and coexists with the member galaxies.

The amount of matter in the intracluster medium can be estimated from the observed X-ray brightness to be comparable to the total mass of all the galaxies in the cluster. Hence at least in



these rich clusters hot plasma is hardly an exotic state of matter but rather one of the dominant forms. Over the past few years X-ray observations of the intracluster gas have yielded good evidence that it originated in the member galaxies; the gas is found to bear the signature of chemical elements such as iron that are thought to be manufactured only in stars. Thus the intracluster medium probably followed rather than preceded the existence of galaxies, and as such it is not a true representative of the hot intergalactic medium. Nevertheless, it stands as a concrete example of the fact that the universe does contain, in at least certain places, large amounts of very hot, X-ray-emitting plasma.

Suppose one accepts for the moment the hypothesis that the diffuse X-ray background can be attributed to bremsstrahlung from a hot intergalactic medium. Apart from resolving an observational curiosity of two decades' standing, would such a plasma have any other interesting properties? The answer is most emphatically yes. In the late 1920's Edwin P. Hubble and his colleagues at the Mount Wilson Observatory established through observations of the spectral features of galaxies that all the galaxies immediately outside the local group are moving away from it; in short, the universe as a whole is expanding. This finding led to the idea that at least the current phase of the evolution of the universe had a discrete beginning in a dense, violent phase; subsequent confirmation of this idea has come from the observation of the microwave background radiation.

Will the present expansion phase last forever? If it does, the universe will be a rather dull place in another 10 or 20 billion years. Not only will all the galaxies have moved far away from any neighboring galaxies but also because

RICH CLUSTER OF GALAXIES serves as proof of the existence on a cosmic scale of an X-ray-emitting plasma: a hot gas of charged particles. The visible-light photograph reproduced in negative form at the top shows the cluster Abell 2151 as viewed by the Palomar Schmidt telescope. The cluster, which is in the constellation Hercules, is about 670 million light-years from the earth. The false-color X-ray image at the bottom, again provided by Chanan, shows the same region of the sky as viewed by the Einstein Observatory. Instead of the individual galaxies seen in the visible-light image, the X-ray image reveals an extended cloud of hot gas, measuring some three million light-years across, entrained by the combined gravitational pull of the assemblage of galaxies. The intracluster medium has an angular diameter of about .25 degree, or half the angular diameter of the full moon. A plasma of this type, estimated from its X-ray emission to have a temperature on the order of 20 million degrees Kelvin, has been found in many similar clusters of galaxies.

stellar evolution is a finite, one-way process that eventually converts galactic matter into faint compact stars such as white dwarfs, neutron stars and possibly black holes, the individual galaxies will eventually lack internal sources of luminosity and will slowly fade from view. This gloomy prediction depends crucially, however, on the assumption that the expansion will continue unabated. The existence of the intergalactic medium can potentially alter the fate of the expansion.

There is a force working to counteract the universal expansion, namely the gravitational attraction of every atom in the universe for every other atom. If the self-gravitation of the universe is sufficiently strong, the expansion will eventually stop and reverse into a gravitationally driven contraction. How can one resolve this important question? Observationally astronomers today know only what Hubble realized half a century ago: the current phase of the universe is one of expansion. Remarkably, a prediction of its future fate—continued expansion or eventual contraction—is in itself not a complicated procedure. Because the magnitude of the gravitational force depends only on the amount of matter involved and the distance between the matter, and the expansion rate and the scale of today's universe can be derived at least in principle from observations of the expansion of galaxies, the question of whether or not the expansion will reverse is in fact equivalent to a calculation of the total amount of matter in the universe.

More rigorously, the mass density of the universe (that is, the amount of matter in any given small volume) is the quantity that needs to be estimated. Above a certain threshold, known as the critical density, the expansion will eventually cease and contraction will begin; alternatively, if the average mass density in the universe falls below the critical density, the expansion will continue forever. The best current estimate of the value of the critical density is exceedingly small, partly because the volume of the current universe is so large; the value is equivalent to an average of about one hydrogen atom per cubic meter throughout the universe.

Apart from the possible existence of the intergalactic medium, most of the matter in the universe is thought to exist within galaxies, in the form of stars and, to a lesser extent, gas and dust. Estimating the mass of the galaxies is a complex and indirect process, but the result in the context of this discussion is so striking that the inaccuracies are probably not critical. The conclusion is that galaxies account for only 1 or 2 percent of the critical density. Thus even allowing for uncertainties in the tally of galactic material, it seems unlikely that normal galaxies provide enough matter to stop the expansion.

The existence of the diffuse X-ray background and its possible origin in an intergalactic plasma, however, demand that one reconsider this calculation. If the mass density in the intergalactic medium exceeds the average mass density in the universe attributable to galaxies by a factor of 50 or 100, the universe would indeed have its critical density, and the prediction of the future evolution of the universe would change drastically. Furthermore, one would have discovered the bizarre fact that astronomers had previously been neglecting as much as 99 percent of the matter in the universe; the familiar, luminous galaxies and their contents would be relegated to the role of a trace impurity in the intergalactic medium!

If the X-ray background is due to the intergalactic medium, one can readily estimate the mass density of the medium by measuring the observed intensity of the background radiation (that is, how much X-radiation arrives at the earth each second from a given direction) and then calculating theoretically how many atoms per volume of space are required to create this much radiation through the bremsstrahlung process. The result of such a calculation proves to be tantalizing. If the X-ray background is due to bremsstrahlung from a hot plasma, then depending on the exact values of several rather poorly measured quantities the inferred density of hot intergalactic plasma is comparable to the critical density of the universe.

Has the discovery of the X-ray background therefore changed ideas about the future expansion of the universe? Unfortunately the background radiation is evidence, but by no means proof, of the existence of a critical density of intergalactic matter. There are three major uncertainties in this chain of argument. First, the value of the critical density depends sensitively on the expansion rate of the universe: the faster galaxies recede from one another, the more matter is needed to restrain the expansion. Estimates of this rate, embodied in the quantity known as the Hubble constant, are derived from a complex sequence of astronomical measurements subject to various uncertainties. Indeed, since Hubble's time the prevailing estimate of this constant has changed considerably. Only for a highly specific (and today somewhat unpopular) value of the Hubble constant does the inferred amount of the intergalactic medium provide the critical density of matter.

A second uncertainty is that the foregoing determination of the density of the plasma based on the observed intensity of X rays calls for a knowledge of the average lumpiness of the emitting regions, because smooth and lumpy mediums emit radiation with slightly different efficiencies. Only if the intergalactic medium is highly uniform is its inferred density equal to the critical density. Although no direct measurement of the degree of lumpiness exists, a smooth medium seems unlikely on theoretical grounds: most constituents of the universe, for example the luminous matter in galaxies, are distributed in a highly nonuniform fashion.

A third uncertainty in the conclusion that a hot intergalactic medium binds the universe is perhaps the most profound one of all: there is no direct proof that the observed diffuse X-ray background is indeed due to such a hot plasma, as opposed to some entirely unrelated mechanism. Unfortunately the X-ray emission bears no unambiguous signature of its origin as bremsstrahlung from a dilute gas, and there are other familiar laboratory processes that can also generate X-ray emission. An explanation of the X-ray background in terms of a hot plasma is consistent with current observations, and perhaps even elegant in its simplicity, but is it unique? Given the cosmological importance of the existence of the plasma, at least if it is near the critical density, one certainly should demand more than a consistent explanation. For example, it would be comforting to know there are no other existing explanations of the X-ray background that are also consistent with the data.

That, however, is not the case. There is an entirely different explanation of the diffuse X-ray background that is completely independent of the possible existence of the intergalactic medium. The competing explanation of the background has been receiving increasing attention among workers in the field in the past few years. The observation that the X-ray background is highly isotropic is a statement not only of a characteristic of the source of the radiation but also of the sophistication of the available instrumentation. Consider the possibility of a large number of discrete objects each of which emits X rays. If the objects were far enough from the observer and sufficiently numerous so that every region of sky held a large number of them, then the observer might misinterpret these discrete sources as a hazy, uniform glow. No matter how accurate a given instrument may be, there would be some hypothetical density of the individual objects in the sky such that the instrument could not resolve them as discrete sources and so determine that an ensemble of points is under observation rather than a genuinely diffuse medium.

Is the diffuse X-ray background really from a diffuse source? Or does it originate from a large number of distant but distinct emitters of X rays, quite unrelated to a dilute intergalactic gas? An ini-

tial step toward settling this issue is obviously to ask whether there is any known constituent of the universe that is both very common and a powerful source of X rays. Galaxies, the basic building blocks of the universe, do not satisfy the requirement. Although they are quite numerous, they simply do not emit enough X-radiation to add up to the observed intensity of the X-ray background.

This conclusion in turn is a consequence of the original conclusion arrived at after the first observation of X rays from the sun: normal stars do not emit a substantial fraction of their energy as X rays. Therefore the sum of the X-ray emission from all the stars in a galaxy is still feeble compared with its visible-light output. I have mentioned a second constituent of the universe that is a powerful source of X rays: the gas in rich clusters of galaxies. This source seems at first to be a promising possibility, since clusters of galaxies are relatively common. Recent observations of the X-ray emission from nearby clusters, together with an estimate of their density elsewhere in the universe, however, tend to indicate that clusters can contribute at most a few percent of the diffuse X-ray background.

In the past three years X-ray observations from the orbiting satellite known as the Einstein Observatory have revealed another class of objects as a strong candidate for the source of the X-ray background. The data show that quasi-stellar objects, or quasars, are strong sources of X-ray emission. The exact nature of quasars is still an enigma. On optical photographs they appear as points, indistinguishable from normal stars in the galaxy. Two decades ago, however, it was realized that unlike normal stars many of these seemingly ordinary starlike objects are intense sources of radio waves.

Detailed studies of the optical emission from quasars prompted by this observation soon revealed that they are receding from the earth at enormous velocities, in some cases at more than 90 percent of the speed of light. If one interprets the recession velocities as being due to the general expansion of the universe, which also causes the recession of external galaxies (a point accepted by most workers in the field but not by all), then one is forced to conclude that quasars are not stars in our galaxy but are the most distant observable objects in the universe. The most extreme examples are far beyond the most distant known galaxies, and they are emitting light that began its journey shortly after the big bang. In spite of the prodigious distances of these objects, many are bright enough to be viewed with an amateur's backyard telescope. Thus the intrinsic luminosity of quasars is enormous, far beyond that of normal galaxies, and their energy-generating mechanism is unlikely to resemble the thermonuclear fusion that powers normal stars.

An examination of carefully selected samples of faint objects on photographic plates has revealed that although few quasars are close to our galaxy, they are a fairly common constituent of the very distant universe. Apparently some evolutionary effect over the past few billion years has largely extinguished the quasar phenomenon.

Before the launching of the Einstein Observatory a handful of the closer quasars were known to be exceedingly powerful sources of X-radiation, based on observations from rockets and earlier satellites. In the most extreme cases a single quasar is known to generate 1,000 times as much energy at X-ray wavelengths as the sum of the visible-light output of all 100 billion stars in our galaxy! Yet because of the vast distances to most quasars even this staggering output of energy is so feeble by the time it reaches the earth that the majority of quasars were not accessible to X-ray observation before the launching of the Einstein Observatory.

The Einstein satellite detected X rays with a sensitivity that exceeded previous experiments by a factor of about 1,000. Furthermore, in contrast to previous detectors of the Geiger-counter type, which essentially recorded just crude X-ray intensities, the Einstein detector carried a novel focusing system that was capable of forming X-ray images, which were then radioed to the ground in digital form. Over the satellite's two-year lifetime it was ordered to obtain X-ray images of a large number of interesting objects, including several hundred quasars. Most of the quasars observed in this way were found to be powerful sources of X-radiation.

Given these X-ray observations of a limited sample of quasars, together with information from the more complete optical studies on the total number of quasars per area of sky, it is possible in principle to estimate the contribution of quasars to the diffuse X-ray background. Such a calculation yields the fascinating conclusion that a large fraction, and possibly all, of the X-ray background is in fact due to distant, individually invisible quasars. As with any survey that relies on a small fraction of a population to extrapolate to the properties of a vastly larger group, however, there are serious uncertainties in it.

In the case of the Einstein Observatory survey of X-ray-emitting quasars the difficulties are compounded. First, the optical data relied on to infer the number of quasars per unit area of sky are still somewhat controversial, and these data are crucial to link the X-ray emissivity of the small sample of objects observed to an estimate of the total contribution of the ensemble of objects. Second, as astronomers have recognized since the discovery of quasars, these strange objects seem to evolve in their properties over the lifetime of the universe. In fact, the X-ray data also provide tentative evidence for the evolution of the X-ray properties of quasars with time; according to workers at the Center for Astrophysics of the Harvard College Observatory and the Smithsonian Astrophysical Observatory, the ratio of the X-ray output to the visible-light output of the quasars they observed seems to depend on one or more other properties, including perhaps how far the object is from our galaxy.

Evidence for such an evolution in X-ray properties is dismaying, particularly because any contribution quasars make to the X-ray background must originate as a mixture of emissions from quasars throughout the universe at a variety of different evolutionary stages. Therefore the exact nature of the evolutionary process must be precisely understood to properly compute the quasars' contribution to the background. (The problem is analogous to that of computing the compound interest on a long-term bank account. If the initial amount of funds and the interest rate are known precisely, it is easy to predict the balance at any given time. If one assumes a slightly different interest rate, however, the predicted balance will be quite different.)

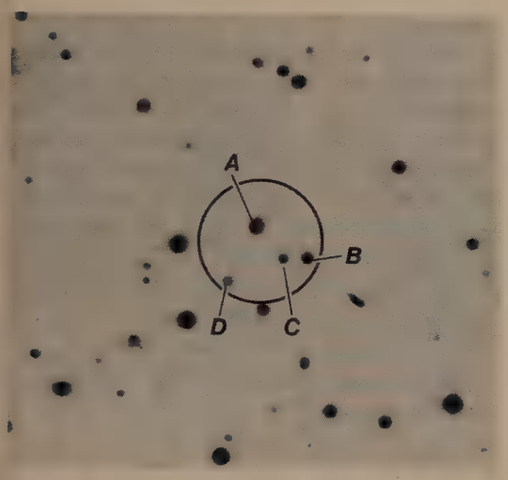
The final problem with the current data is one that is familiar to all statisticians and census takers: How normal is the small fraction of the population selected for the survey? The Einstein Observatory was commanded to point at quasars discovered over the past few years because they are extremely luminous in either their visible-light output or their radio output; these objects represent the large majority of the previously known quasars. The question is: Will such objects yield an unbiased survey of the X-ray properties of quasars?

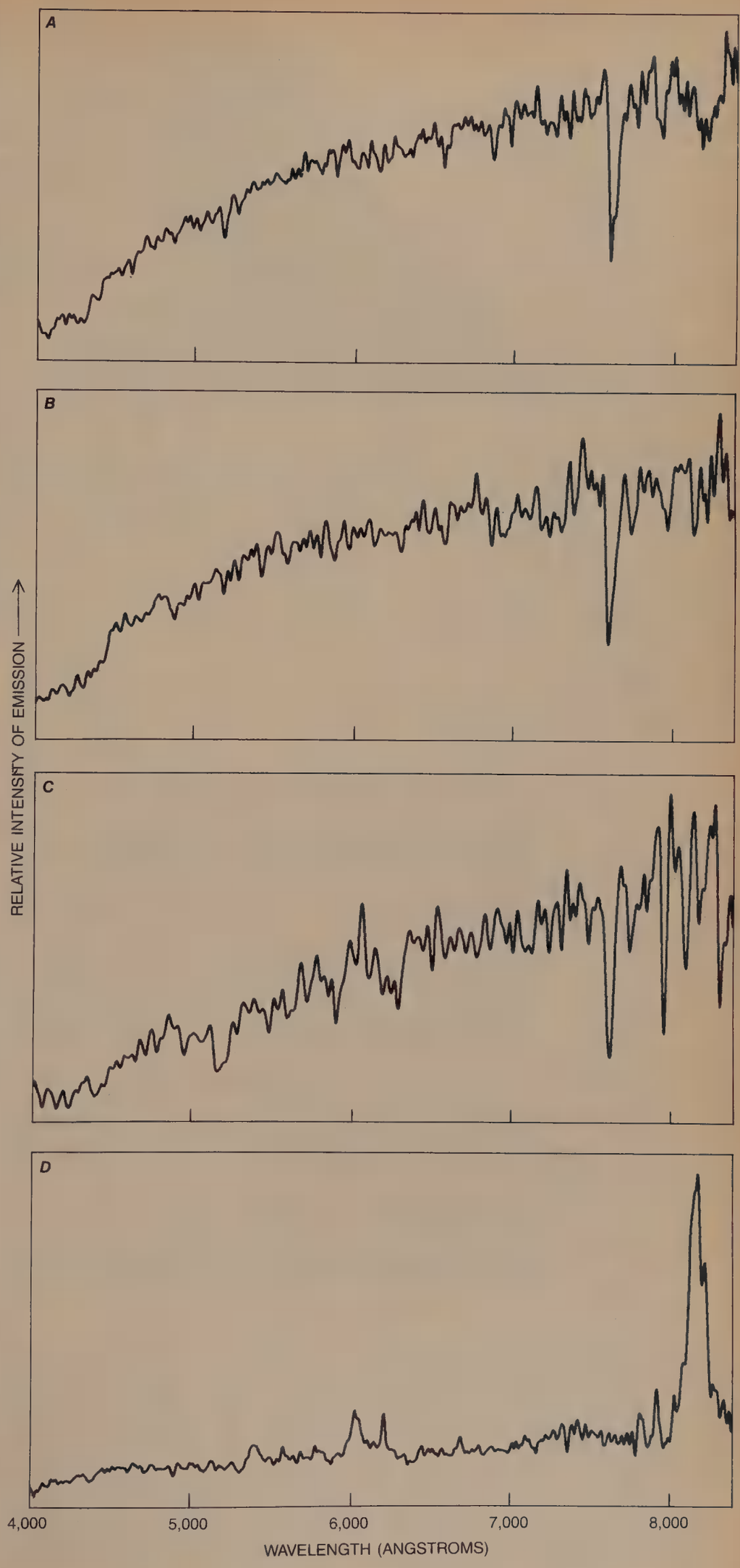
At the beginning of the Einstein program there was obviously no alternative to the practice of observing the quasars from existing catalogues, but the remarkable sensitivity of the Einstein X-ray detectors has rather unexpectedly provided an alternative. A typical observation consisted of orienting the satellite so that the astronomical object of interest was in the center of the focused X-ray image prior to the start of the exposure. As with any camera some surrounding area unrelated to the subject of interest was included in the image. An examination of the Einstein X-ray images reveals that a large fraction of these records show one or more additional faint X-ray sources at the edges of the image, unrelated to the original subject of the exposure. These newly dis-

covered X-ray sources have come to be called the serendipitous Einstein sources because of the accidental nature of their discovery.

The longer the X-ray exposure is, the more serendipitous sources tend to accumulate on a given Einstein image. The finding provides strong evidence that if point sources of any type are a significant contributor to the diffuse X-ray background, then the serendipitous sources are probably representative of this population. In other words, the serendipitous sources appear in the data unpredictably and without the intervention or advance knowledge of the observer; they are the pollster's dream of a truly unbiased sample.

What is the nature of the serendipitous sources? In particular, do they cor-





SERENDIPITOUS DISCOVERY of a previously uncatalogued X-ray-emitting quasar is depicted. In the course of a study of the X-ray emission from a binary stellar system in the constellation Aquarius a second X-ray source was noted nearby on an Einstein Observatory X-ray image. On the basis of the X-ray data the position of the new source could be established only approximately. On a visible-light photograph of the same region of the sky (*above*), made with the Palomar Schmidt telescope, a circle is drawn around the region where the X-ray source appeared to be. Four starlike objects are visible within the circle, all of which were treated as candidates for identification with the X-ray source. Spectra of each of these extremely faint objects were obtained by the author and his colleagues with the three-meter Shane reflector of the Lick Observatory (*curves at right*). The first three spectra have prominent dips where cool atoms in the outer layers of the object have apparently absorbed radiation from the interior. Such spectral absorption features are characteristic of normal stars, indicating that these three objects are probably not related to the X-ray source. The fourth spectrum has a completely different appearance: prominent peaks that correspond to excessive emission at certain wavelengths. This is the telltale signature of a quasar, in this case one that can be shown from its spectral features to be some 3.8 billion light-years away. Designated 2216–043 for its celestial coordinates, the newly discovered quasar is the optical counterpart of the serendipitous X-ray source.

respond to other objects already well studied, for example by optical methods, or are they some entirely different constituent of the universe? The question is not easy to answer because the limited focusing ability of X-ray optics results in images that are significantly blurred by the standards of optical astronomy. Just as it is difficult to measure precisely the characteristics of a blurred image on a visible-light photograph, so it is hard to assign an exact position in the sky to each of the serendipitous X-ray sources. Instead one establishes a small region, about a thirtieth of a degree across, from within which one can estimate that the X rays emanate. This precision is not sufficient to identify the object uniquely with any known visible source; a visible-light photograph of the same region of the sky where the X rays originate will typically show between four and six faint objects, any one of which (or none of which) could correspond to the X-ray source.

Further observations are clearly necessary to reveal the nature of the serendipitous X-ray sources and to determine if they have any bearing on the problem of the diffuse X-ray background. Several groups around the world have conducted such observations, including one consisting of Gary A. Chanan of Columbia University, Ronald A. Downes of the University of California at Los Angeles and me. Our technique has been to examine ordinary visible-light photographs of the small region of the sky that corresponds to the region of positional uncertainty of each of the serendipitous sources. As I have noted, this procedure generally yields a handful of faint objects that are candidates for the visible-light counterpart of the X-ray source. We then examine these visible objects in detail, searching for some anomalous property. (Bear in mind that normal stars are rarely detectable X-ray sources.) Because the candidates are typically 100,000 times too faint to be seen with the unaided eye, large telescopes are needed for the purpose. We worked primarily with the 2.1-meter reflector at the Kitt Peak National Observatory, the three-meter Shane telescope at the Lick Observatory and the four-meter reflector at the Cerro Tololo Inter-American Observatory in Chile.

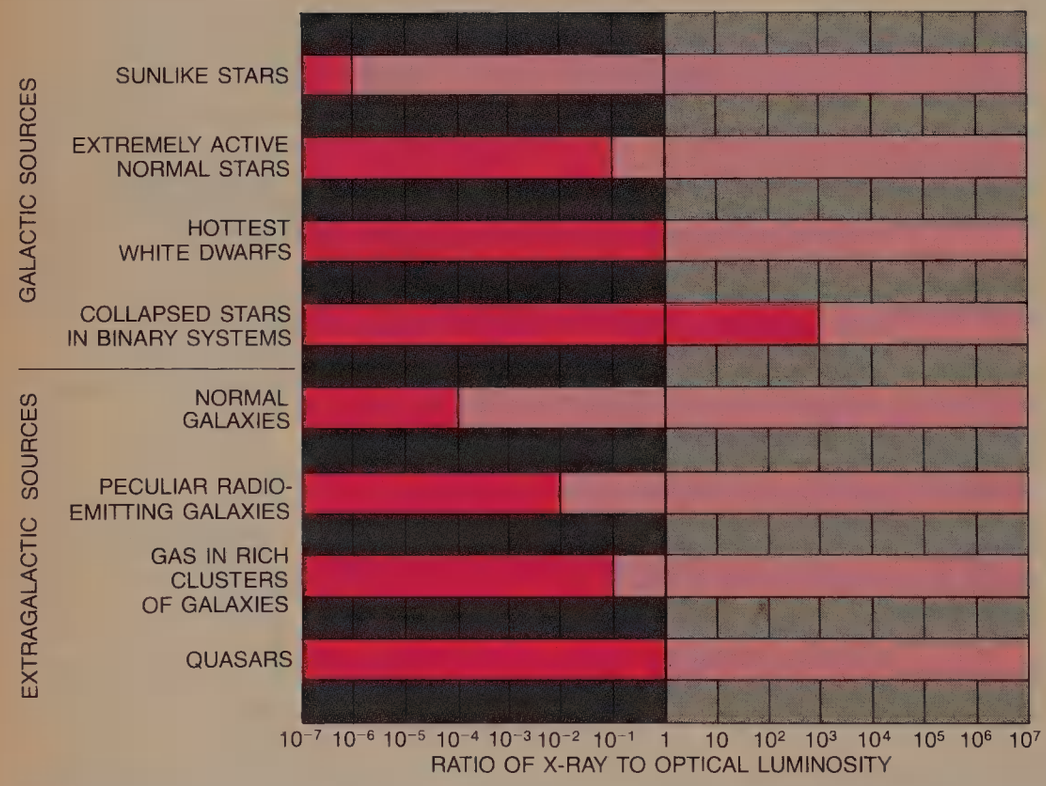
Our technique is based on spectroscopy, the division of radiation into its component wavelengths. The spectrograms of normal stars have gaps at characteristic wavelengths, where cool atoms in the upper layers of the star have absorbed light before it could reach the detector. Typically the spectra we obtained of almost every stellar object near the location of the serendipitous X-ray source showed this familiar pattern; apparently all these objects are normal, comparatively nearby stars, completely unrelated to the source of the X rays.

In most instances, however, we have found another object in the region that has a radically different spectrum, with a discrete set of wavelengths at which there is an excess of radiation. This pattern of emission features, precisely the opposite of the absorption features found in normal stars, indicates a highly unusual object, evidently one with copious amounts of hot gas. Furthermore, the precise wavelengths where the emission features appear correspond to those of familiar atoms such as hydrogen, helium and carbon, but with a large, systematic displacement in wavelength toward the red end of the spectrum. The red shift indicates that the source of the light is receding from the observer. This peculiar combination of spectral characteristics—strong, red-shifted emission features associated with common chemical elements—is the signature of a quasar. In short, what we, along with several other groups working on the problem, have found is that a large fraction of the serendipitous X-ray sources are actually previously undiscovered quasars.

The finding obviously has profound implications for the question of the origin of the diffuse X-ray background, and in particular for those interpretations in which the background radiation is attributed to an ensemble of distant, pointlike sources. If one peers through a fog and can just barely discern the nature of the brightest of a large number of distant objects that fade into a blur, it seems a reasonable assumption that similar objects constitute a significant fraction of the more distant, individually unobservable population. Similarly, the identification of a large number of the serendipitous X-ray sources with quasars (about 150 such cases are currently known) strongly hints that quasars are responsible for a significant fraction of the background radiation.

How large a fraction? Are quasars the predominant contributor to the background, thus solving the 20-year-old problem of the origin of this radiation, or are they merely a curiosity, a trace contaminant of a background due instead to a hot intergalactic plasma? Because there is no estimate independent of the intensity of the X-ray background of how much hot intergalactic medium exists (if any), there is a real possibility that both sources contribute to the observed X rays.

The properties of the ensemble of serendipitously discovered quasars are the key to the resolution of this issue. A study of the properties of the objects in



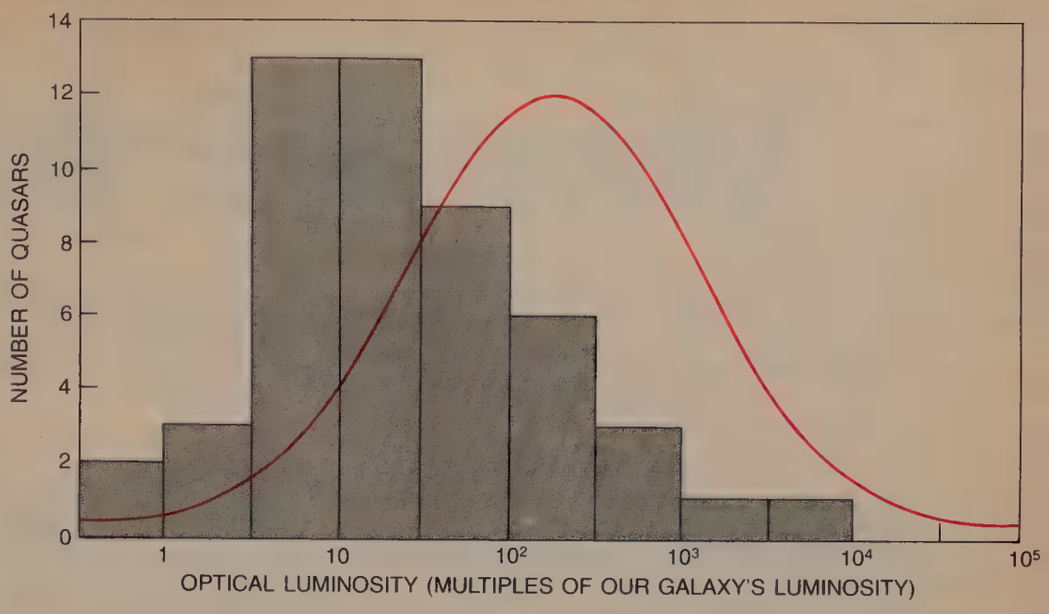
MOST ASTRONOMICAL OBJECTS generate only a small fraction of their total output of radiation at X-ray wavelengths. If the X-ray background is attributed to a large number of very distant individual sources, the objects must be both powerful X-ray emitters and fairly common in order to account for the observed isotropy. As these bars indicate, this condition is difficult to meet: the powerful X-ray emitters also tend to be rather exotic objects. Individual stars whose ratio of X rays to visible light is 1 or more are so rare that the sum of the X-ray emission from a normal galaxy is typically only a 10,000th (10^{-4}) of its visible-light output. Hence normal galaxies, although they are numerous, cannot contribute appreciably to the diffuse X-ray background. The hot plasma trapped in rich clusters of galaxies does generate an abundance of X rays, but such clusters are too rare to account for more than a few percent of the X-ray background. Quasars come closest to the required combination of properties: their X-ray output is prodigious, and they are particularly numerous at great distances from earth.

this sample, the first quasars discovered by their X-ray emission rather than their optical or radio emission, should provide a new and perhaps less biased insight into the X-ray properties of these enigmatic objects. Indeed, a comparison of the properties of the X-ray-selected sample with previous samples of optically selected or radio-selected quasars does reveal interesting differences. The intrinsic amount of light energy emitted by the X-ray-selected group seems to be systematically and significantly less, by a factor of about 10, than the corresponding quantity for the ensemble of previously known quasars.

A precise interpretation of the difference, however, is at the moment unclear. Some workers argue that the ensemble of all quasars has a broad range of possible X-ray and visible-light luminosities. Optical surveys would then find the most conspicuous visible-light objects with rather normal X-ray characteristics; the X-ray surveys would find exactly the opposite. Then a comparison of the two samples might well reveal significantly different properties, as are in fact observed. Alternatively, certain of the evolutionary effects cited above could preferentially suppress the X-ray emission of a subset of quasars, again creating different properties for X-ray-selected and optically selected samples. Worse yet, both effects may be important simultaneously.

The possible evolutionary effects are particularly troublesome. Although any isotropic radiation must be due to the sum of constituents at various distances from the observer, the dominant contribution can be shown to be made by the most distant objects. If the properties of the objects evolve and change progressively with distance, however, these most distant, extremely faint objects, for which the poorest data are available, can be very different from the closer, brighter objects. For the latter group the data may be good, but these data are not typical of the population of interest. This effect is almost certainly important at some level for the X-ray-emitting quasars. Even the extraordinary sensitivity of the Einstein detector was not sufficient to examine the objects that must be responsible for the bulk of the X-ray background, if indeed quasars are the main source of the background. Therefore extrapolations of the properties of the X-ray-selected quasars are needed to assess the total contribution of these objects to the diffuse radiation. At the moment the accuracy of the extrapolations is not known. Indeed, depending on the precise nature of the procedure, one can conclude that quasars are either a dominant component of the background or a negligible one.

Although the serendipitously discovered X-ray-emitting quasars provide new insights into the contribution of quasars to the background radiation, the exact interpretation of these insights is clearly complex and at the moment ambiguous. As often happens in observational astronomy, a large increase in sensitivity has brought not only new answers but also many new questions. Are there any independent clues to aid in separating the contribution of quasars to the X-ray background from that of the intergalactic plasma? One possibility lies in determining the spectrum of the observed radiation: How much X-radiation is detected at each wavelength compared with neighboring wavelengths? This technique is potentially a powerful diagnostic tool, because the different fundamental physical processes that generate X rays often result in different, recognizable spectra.



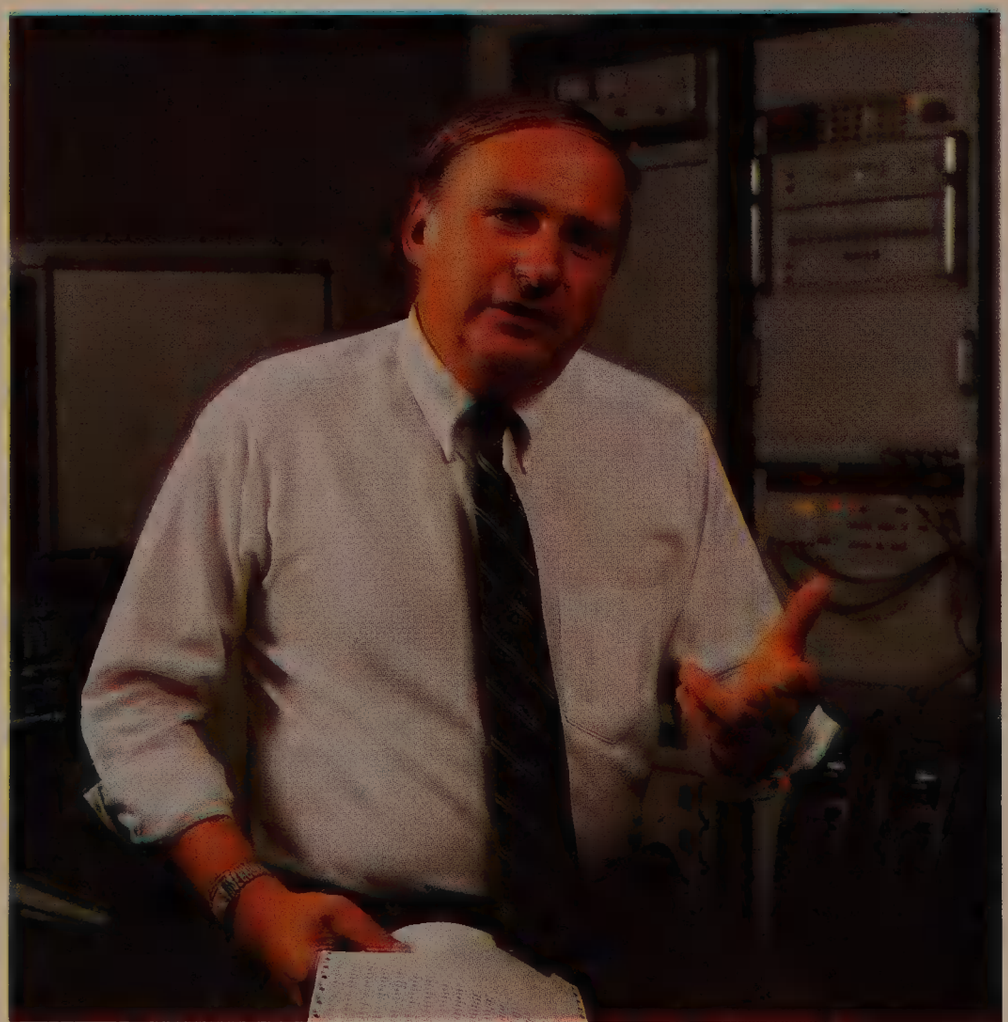
FIFTY-ONE QUASARS discovered by the author and his colleagues as the optical counterparts of serendipitous X-ray sources appearing in the images recorded by the Einstein Observatory are plotted in this bar chart according to their visible-light output. Prior to this work virtually all the known quasars were discovered because of their unusual optical or radio emission rather than their X-ray emission. The colored curve indicates the optical-luminosity distribution for a representative sample of 51 optically selected and radio-selected quasars. The X-ray technique evidently selects quasars with an optical luminosity lower than that of the quasars selected by the earlier methods. The difference in luminosity between the X-ray-selected quasars and the previously discovered quasars is a clue to estimating the X-ray output of a typical quasar and hence to determining the contribution of quasars to the diffuse X-ray background.

The experimental techniques for spectroscopy at X-ray wavelengths are more difficult than those at the wavelengths of visible light. Nevertheless, satellite-borne instruments have determined the spectrum of the X-ray background with increasing accuracy over the past few years. Much of the work in this area has been done by a group at the Goddard Space Flight Center of the National Aeronautics and Space Administration. Workers there have pointed out that the observed spectrum of the background radiation closely resembles the one expected from bremsstrahlung from a hot plasma at a temperature of about 500 million degrees K. This result can be taken as a powerful point in favor of the argument that a hot plasma dominates the background radiation, and that contributions from quasars (or any other type of distant discrete source) are quite small. As usual, however, there are complexities in the interpretation of the result that weaken the conclusion.

One such difficulty is in understanding what spectrum might be expected if an ensemble of quasars, rather than a hot gas, constitutes most of the background. In that case the spectrum would be the sum of the individual spectra of a large number of objects. X-ray spectra have been individually determined for only a few of the brightest quasars, and they do not particularly resemble the observed spectrum of the background. Perhaps the X-ray spectra of distant quasars, however, are different from the spectra of nearby quasars, again because of evolutionary effects. Furthermore, the spectrum of the diffuse background is most accurately determined at a range of X-ray wavelengths slightly different from the wavelength where the Einstein imaging observations were made. If nature were sufficiently mischievous to change the emission mechanism responsible for the bulk of the radiation from one regime or the other, workers determining the spectral properties of the radiation and workers concentrating on the imaging observations might in fact be unknowingly attempting to solve two slightly different problems.

There is one reasonably unambiguous measurement available to help distinguish the contribution to the X-ray background of distant, discrete X-ray sources from that of a hot plasma. As I have mentioned, the longer the exposure obtained by the Einstein Observatory,

What if you chos
as a technica



"At M/A-COM, HP computers increase test speed 12-fold."

M/A-COM, Inc., is a major manu facturer of electronic products for commercial telecommunications and defense markets. Dr. Alan Carlson, Manager of the Automa Test Equipment Group in Burling ton, Mass., says: "We build auto- matic test systems and conduct final test on hundreds of M/A-CO products. To automate, we neede precision instruments plus computer power.

"HP was the only vendor to offer us a complete solution. With 20 HP 1000 computers driving more than 200 HP instruments, we've completely automated several pr duct tests. And we can depend o HP's support teams to keep our entire testing network up and running.

"Now we can test sophisticated products 12 times faster than befo With these kinds of savings, each HP 1000 has paid for itself in less than one year."

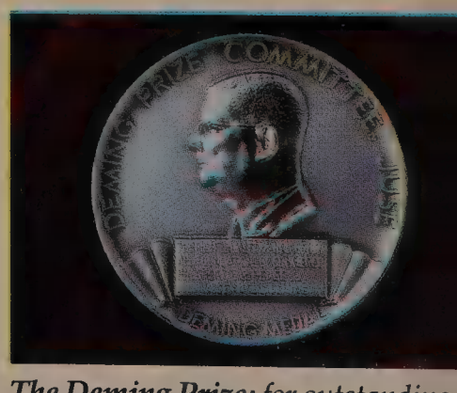
What should you expect from a computer partner?

Quality—from researching customer needs to product development, manufacturing, marketing, after-sales service, and support!

Quality, by HP's definition, encompasses more than the product.

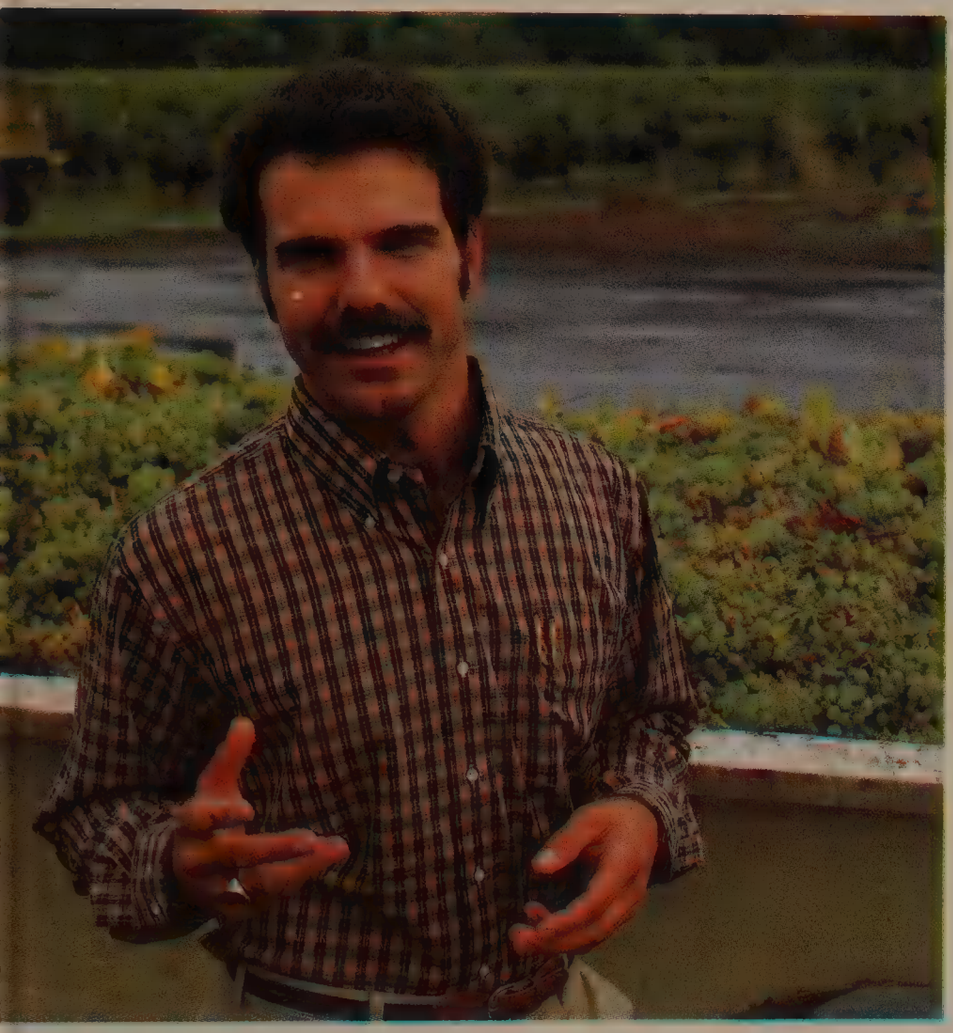
This high quality recently was recognized by the Union of Japanese Scientists & Engineers, who awarded a coveted Deming Prize to Yokagawa-Hewlett-Packard, HP's joint venture company in Japan.

YHP offers the entire spectrum of HP products and in 1981 accounted for Japanese sales of more than $200 million. Nearly 60 percent of the products YHP sells are imported from HP manufacturing operations



The Deming Prize: for outstanding merit in the control of quality.

Hewlett-Packard computer partner?



"At Charles Krug Winery, an HP computer tests temperature 150 times faster than manual methods, improves quality control, and doubles production."

Charles Krug and CK Mondavi premium wines are produced at the Charles Krug Winery in St. Helena, California. The winery uses an HP 1000 computer system to monitor and control the temperature of its fermentation tanks.

Treasurer Peter Mondavi, Jr. explains: "The HP 1000 tests the temperature of each tank 150 times faster than manual methods. And because it automatically adjusts fluctuating temperatures we know each bottle of wine will meet our quality standards.

"Improved temperature control has enabled us to double our production while maximizing quality control. As a result, we've saved $50,000 in production costs this year alone."

he US and abroad. Conversely, P manufactures and exports ious products to HP customers oughout the world.

e total quality control approach which YHP was awarded the ning Prize is a point of pride oughout HP. Customers in every t of the world should expect *lity* as a fundamental element in mputer partnership with HP.

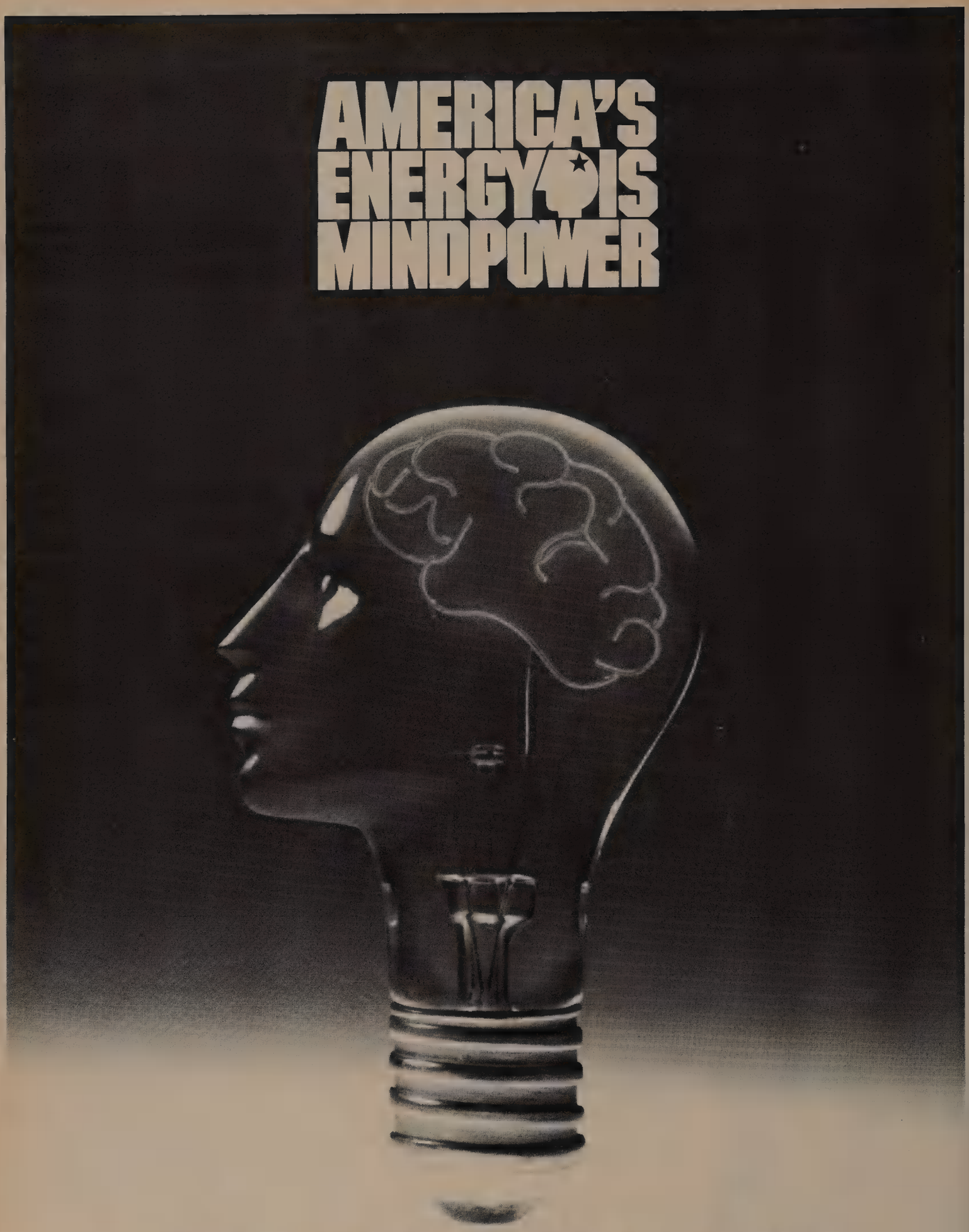
You should consider a working partnership with HP. Now.

For a free copy of our brochure explaining HP's practical, proven approach to meeting your long-term computer and information needs, write to A.P. Oliverio, Senior Vice President, Marketing, Dept. 207, Hewlett-Packard Company, P.O. Box 10301, Palo Alto, CA 94303.

When performance must be measured by results

HEWLETT PACKARD

0002340



MAKE HIGHER EDUCATION YOUR PRIORITY

Support our colleges: They provide our nation's greatest resource.

the greater the number of serendipitous X-ray sources detected. Therefore the fractional contribution of X rays represented by these discrete sources to the total number of X rays in an entire image is certainly a conservative lower estimate of the fraction of the background due to such discrete sources. On the deepest exposure this fraction is about a fourth. Clearly objects such as quasars do make a significant contribution to the background radiation. This measurement, however, is well short of the 50 percent mark, where one could say with confidence that discrete objects are responsible for the dominant contribution to the background radiation, thus essentially resolving the issue.

It should be clear from this discussion that a more powerful X-ray imaging instrument than the ones carried by the Einstein Observatory has the potential of settling the question in a direct and elegant manner, simply by individually resolving enough weak sources to account for half or more of the total emission in any small patch of sky, or alternatively discovering that at very low levels of X-ray brightness the radiation remains truly diffuse and does not break up into separate sources. The requisite technology is available, and a suitably powerful instrument has been proposed to NASA for flight in a satellite. The development of this project, known as the Advanced X-ray Astrophysics Facility (AXAF), would require approximately 10 years once approval for funding was obtained. A committee of astronomers headed by George B. Field of Harvard University recently prepared a list of important priorities for the next decade of astronomical research, and the list ranked AXAF first among its recommendations.

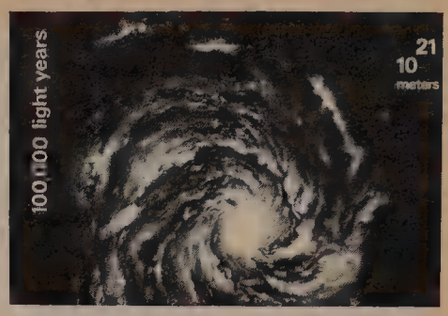
It is unclear at the moment whether further studies of the serendipitously discovered X-ray quasars may make it possible to understand with reasonable accuracy the possible effects of the evolution of quasars. Such an understanding, coupled with more sophisticated optical data on the number of quasars typically found in any small region of the sky, might then lead to a precise calculation of the importance of quasars to the background and by implication to an estimate of the contribution of the hot intergalactic plasma. The necessary higher-sensitivity optical data will almost surely be provided by the Space Telescope, which is now scheduled to be launched into orbit in 1985. Either at that point, or perhaps in the early 1990's with the launching of the AXAF satellite, it may finally be possible to know with confidence whether the diffuse X-ray background and quasars, both discovered independently 20 years ago, are actually two separate manifestations of one phenomenon.

SCIENTIFIC AMERICAN

is now available to the blind and physically handicapped on cassette tapes.

All inquiries should be made directly to RECORDED PERIODICALS, Division of Volunteer Services for the Blind, 919 Walnut Street, 8th Floor, Philadelphia, PA 19107.

ONLY the blind or handicapped should apply for this service. There is a nominal charge.



POWERS OF TEN
By Charles and Ray Eames

This famous film by Charles and Ray Eames transports you out to the edges of the universe and in to the micro-world of cells, molecules and atoms.

The film inspired the book POWERS OF TEN, the premier volume of the SCIENTIFIC AMERICAN LIBRARY.

POWERS OF TEN—another outstanding film available from PYRAMID FILM & VIDEO.

POWERS OF TEN 9 minutes, color
$225 Purchase: 3/4" U-Matic, 1/2" Beta II, 1/2" VHS, 16mm-Rent: $40 3/4" U-Matic and 16mm only
call toll free: 800-421-2304

PYRAMID FILM & VIDEO
Box 1048, Santa Monica,
California 90406-1048 (213) 828-7577

Learn Spanish the Foreign Service way!

You can be fluent in Latin American Spanish. That's right, *fluent!* You can be speaking with all the confidence of a seasoned traveler in a matter of weeks. As well-versed in Spanish as if you worked for the U.S. Foreign Service Corps.

The Foreign Service Institute has carefully developed, tested, and refined its self-instructional audio cassette Spanish course to train the U.S. State Department's overseas personnel quickly, thoroughly, and effectively. Its clear, easy-to-follow approach and arrangement of material makes this a perfect method for successful self-instruction.

Now the FSI's complete program is available to *you*. We will send you the entire audio cassette series and instruction books to lead you step by step to fluency in weeks. You work when and where you can, at your own pace, following the taped lessons. You progress according to the schedule *you* establish.

The course uses native speakers exclusively so that you hear exactly how each word and sentence should sound. It won't be long before you start speaking and thinking in your new language. **This program has to work—our State Department depends on it!**

MONEY-BACK GUARANTEE: Try the FSI Basic or Intermediate course for three weeks. If you're not completely satisfied, return it for a full refund. **There's no risk and no obligation, so order today.**

SPECIAL OFFER: Purchase any language course from Pinnacle and get the **Toshiba KT-S3 Personal Stereo Cassette Player,** complete with headphones, FM Tuner Pack, and shoulder strap (nationally advertised for $139.00), for **only $89.95.**

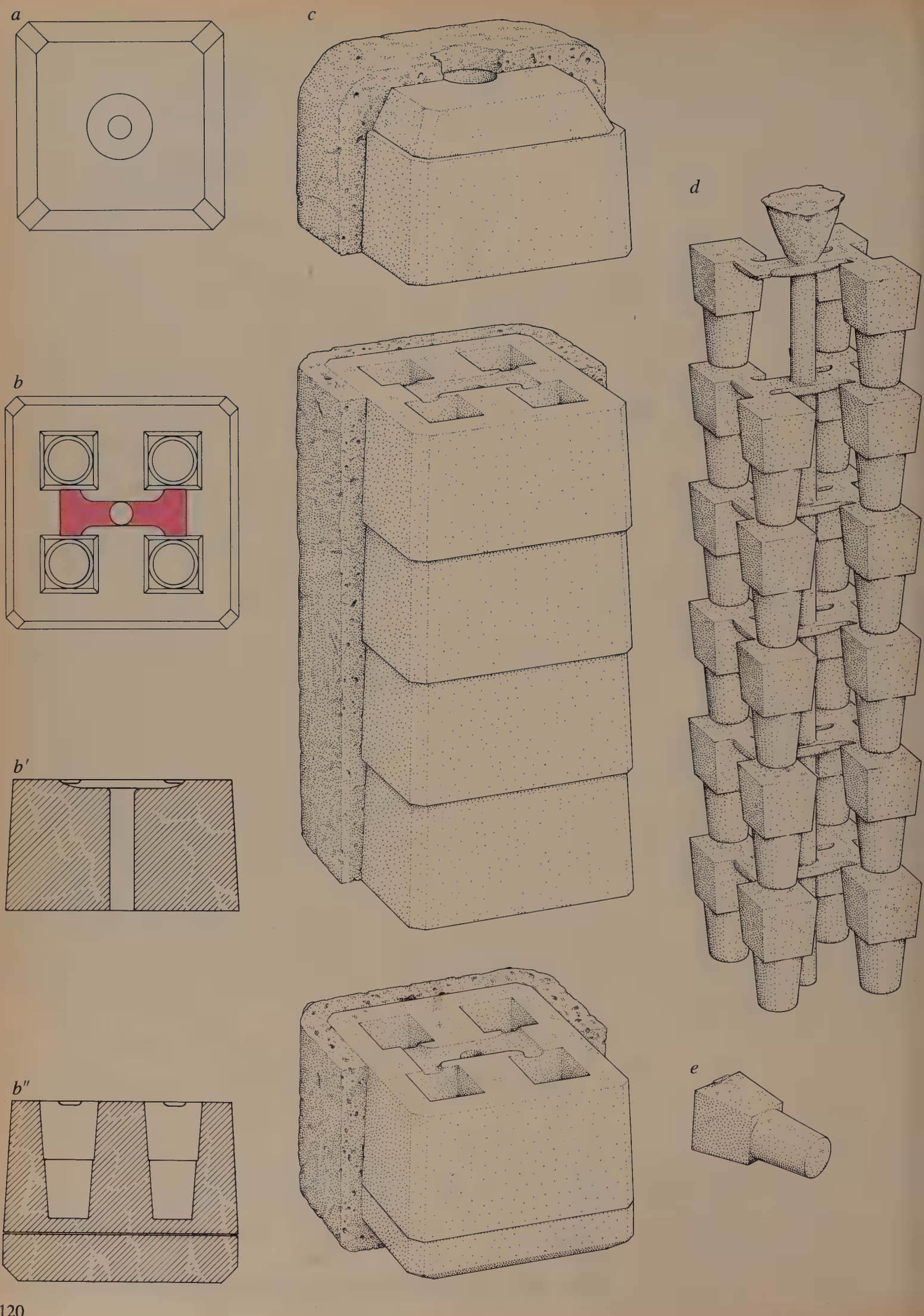
French also available. Write us for details or call (212) 772-9202.

pinnacle **Dept.209, 215 E. 79th St. New York, N.Y. 10021 (212) 772-9202**

TO ORDER BY PHONE—(Credit Card Orders Only Please)—Call us Toll Free: **1-800-528-6050 Ext. 1611** In AZ 1-800-352-0458 Ext. 1611

Please send me the following course(s) shipped in attractive library albums.

☐ **Programmatic Spanish—Vol. 1 Basic:** 11 Cassettes 16 hrs. manual and 464-page text. **$100**
☐ **Programmatic Spanish—Vol. 2 Intermediate:** 8 Cassettes 12 hrs. manual and 614-page text. **$90**
☐ **Both Volumes for Only $180.** ☐ **Toshiba KT-S3. $89.95** plus $4.00 shpg.
☐ Enclosed is my check or MO for $_____ (In NY add sales tax.)
Charge my: ☐ MasterCard ☐ Visa No. _____ Expiration Date
☐ Amex ☐ Diners Club
Signature _____
Name (please print) _____
Address _____
City _____ State _____ Zip _____


a
b
b′
b″
c
d
e

The Mass Production of Iron Castings in Ancient China

By 500 B.C. the Chinese had developed stack casting: a high technology in which multiple castings are made by pouring iron into multiple molds designed to be stacked one on top of another

by Hua Jue-ming

The casting of bronze in China, going back at least to the Xia Dynasty (21st to 16th centuries B.C.), is one of the best-known traditions of ancient technology. Less familiar outside China is a development of the first millennium B.C.: the mass production of iron castings. Chinese foundrymen made coins, harness buckles, vehicle parts and other small articles in huge numbers by the technique of stack casting. The term refers to the fact that individual molds are stacked one on top of another and share a single "gate," or point of entry for the molten metal. The technique made it possible to fill as many as a dozen molds at the same time, thereby economizing on both metal and fuel. Stack casting began in China between 800 and 500 B.C. and is still practiced today.

Recent archaeological investigations have uncovered the remains of stack-casting foundries in four provinces: Henan, Shandong, Shaanxi and Jiangsu. Perhaps the richest site, found almost a decade ago by Lee Jing-hua and Tang Wen-xing of the Henan Provincial Museum, is a kiln west of the city of Wenxian in Henan. During the Eastern Han Dynasty (A.D. 24–220) the kiln was a key part of a major stack-casting foundry. Excavations at the site revealed that the kiln still held several hundred intact sets of molds, some of them designed to produce more than 80 castings at a time. Many of the ancient molds were so well preserved that the investigators have been able to use them to create objects identical with those first produced in cast iron nearly 2,000 years ago.

Before describing what has been learned at the Wenxian kiln site it will be useful to acquaint the reader with the basic technology of metal casting and the steps involved in it: first the making of the molds and then the melting, pouring and freezing (that is, the solidification) of the metal. Most of the work in casting technology has to do with the making of the molds. Throughout antiquity and into modern times the ideal mold material has been clay. It is abundant in all parts of the world, and little more than screening and washing to rid it of larger particles and foreign matter is needed to prepare it for the making of molds.

When dry clay is mixed with water, the material's exceedingly fine platelike particles become coated with a film that allows the plates to slide past one another on the application of very little force. At the same time the particles cannot easily be separated from one another because of the surface tension of the water film. Depending on the proportion of water to clay the wet mixture can vary from a stiff paste to a thin slurry. When drying removes the water, the clay reverts to a solid of modest strength. That strength can be greatly enhanced by firing (baking in a kiln); the heat initiates processes that bond between the clay particles. Like most other peoples of the world, the Chinese were familiar with these properties of clay as a result of making pottery for millenniums before the rise of metallurgy.

Clay pastes and slurries both play a role in mold making. A component of a mold can be made out of a paste and then formed into the desired shape by pressing into the soft surface of the clay a model made of wood or metal (a "pattern"). Conversely, a pattern having the desired shape can be repeatedly painted with a slurry and allowed to dry until the pattern can be lifted away from what is by then a thick clay shell that perfectly duplicates the pattern's detail. Either way the pattern can serve repeatedly for making a very large number of identical mold segments.

The mold segments must next be dried (to evaporate most of the water the clay still holds) and then fired. This not only increases the strength of the mold but also rids the clay of any remaining water and eliminates the hazard of steam bubbles' causing surface defects when the metal is poured or otherwise distorting the form of the casting. Firing the molds can make another contribution to casting. If, as was true at the Wenxian foundry, the stacked molds are taken from the oven while they are still quite hot and the molten cast iron is poured into the hot stack, the freezing process is retarded, allowing the fluid metal to fill all the mold interstices.

This retardation of freezing was a factor of considerable importance to the workmen at the Wenxian foundry because their stack molds had particularly narrow "runners": the connecting slots that conducted liquid iron from the central sprue, or vertical channel, to the various mold cavities. Premature freezing of the iron in these narrow slots

SIX-MOLD STACK for the production of 24 key wedges from a single pouring of high-carbon "white" cast iron appears in the illustration on the opposite page. At the left, shown in plan view only, is the "gate" (*a*) at the top of the stack, which provides a central pouring point for the molten metal. Below the gate, also in plan view, is a single mold with four casting cavities (*b*) and "runners" (*color*) leading to them from the central "sprue." The elevation view below it, (*b'*), running through the midline, shows the runners and the continuation of the sprue. The elevation view at the bottom (*b''*) runs through two of the four casting cavities. At the center (*c*) is the stack assembly shown with the top and bottom components exploded. The dark outer coating, half cut away, is a fired layer of straw and mud that made it possible to handle the six molds as a unit. At the right (*d*) is the casting as it appeared after the iron had solidified and the clay molds had been broken away. One wedge, runner removed, appears below the casting (*e*).

would have resulted in incomplete castings. At the same time the mold designers could not simply eliminate the problem by making the runners larger. The reason was that the metal they used was not the ordinary gray cast iron of the Western world but white cast iron, a form of the metal that is very much harder. Indeed, white cast iron is so hard that the labor of breaking thick runners off the finished castings and then grinding any rough areas smooth would have seriously cut the rate of production.

Bronze, the first metal cast on a large scale in China, consists mainly of copper. Iron ores are much commoner than copper ores, so that it is not surprising iron became the workhorse metal in China, as it did elsewhere in the world. In Europe and as far east in Asia as India the first iron implements were made by smelting and smithing. The product of the smelting process is a spongy, slag-filled mass of almost pure "bloom" iron. The bloom is converted into usable metal by hot hammering that squeezes most of the viscous slag out of it. Further heating in a forge and hammering on an anvil yields finished iron objects such as tools, weapons and ornaments.

In China iron production followed a different track. The Chinese, using more fuel and pumping more air into their smelters than the metalworkers of Europe and West Asia did, attained conditions that completely melted the iron ore, yielding a metal free of slag and rich in carbon. The carbon was important because the greater its concentration in the iron, the lower the metal's melting point. Pure iron does not melt below 1,535 degrees Celsius, a temperature that was not attained in furnaces on a practical basis until the 19th century. When iron contains 4.3 percent of carbon, however, it becomes completely liquid at 1,130 degrees C., a temperature only about 80 degrees higher than the melting point of bronze. Once this property of white cast iron was discovered in China the Chinese foundrymen, already masters of casting bronze, recognized they could do the same with the carbon-rich metal, thereby utilizing a much more abundant metallurgical resource.

By the middle of the first millennium B.C. cast iron was common in China. Weapons, ax and adz blades, hammerheads, hoe blades, spade tips and plow tips were being cast in iron; so were decorative items such as belt buckles and rings, household articles such as pots and braziers and equipage such as horse bits and the axle fittings of chariot wheels. A period of more than a millennium, from about 300 B.C. to perhaps A.D. 1100, saw a great growth of all technology in China. It was probably also a period of population growth. As one factor in this dual growth consider the plow. It harnesses human and animal power not only for routine cultivation but also for breaking new ground. For example, cast-iron plow tips broke the virgin prairie in the 19th-century march of agriculture across North America. In China as elsewhere wood plows preceded iron-tipped ones. It is not hard to imagine that with the availability of cast-iron plow tips toward the end of the first millennium B.C. there came a revolution in the increase of arable land, in food production and therefore in food for an expanding Chinese population. With the growth of agriculture and handicrafts came the need for transportation, and the demand for harness and vehicle parts was great. Such was the setting for the work at Wenxian.

The Wenxian kiln was a part of a specialized foundry mainly producing iron castings for horse-drawn vehicles. The 500 sets of molds found at Wenxian could turn out 16 different kinds of castings in a range of 36 sizes. The castings included horse bits, harness rings, harness buckles and vehicle fittings such as bearings and axle end pieces. Among the molds for other kinds of castings were those for the weights of the steelyard, a type of balance found in many markets even today.

The uniformity of the molds suggests they were formed in master mold boxes made of metal, although no such masters were found at the site. One advantage of having such master patterns, in addition to the uniformity it ensures, is that any two halves of the mold made from the master can be mated to form a complete pair. This appears to have been particularly useful for the manufacture of buckles; some of the stack molds for these objects, with 14 levels and six buckle impressions per level, yielded 84 castings in a single pouring. An even greater advantage is that with many master patterns available many mold makers can work at the same time, which greatly increases productivity.

Analysis of vegetable debris at the kiln indicates that the metal masters were dusted with a fine bran before clay paste was packed into them and pounded down to acquire the negative impression. The bran facilitated the clean parting of the clay mold from the master; the molds were then left to dry for a week or so in a cellar or some other shaded place.

When the foundrymen had accumu-

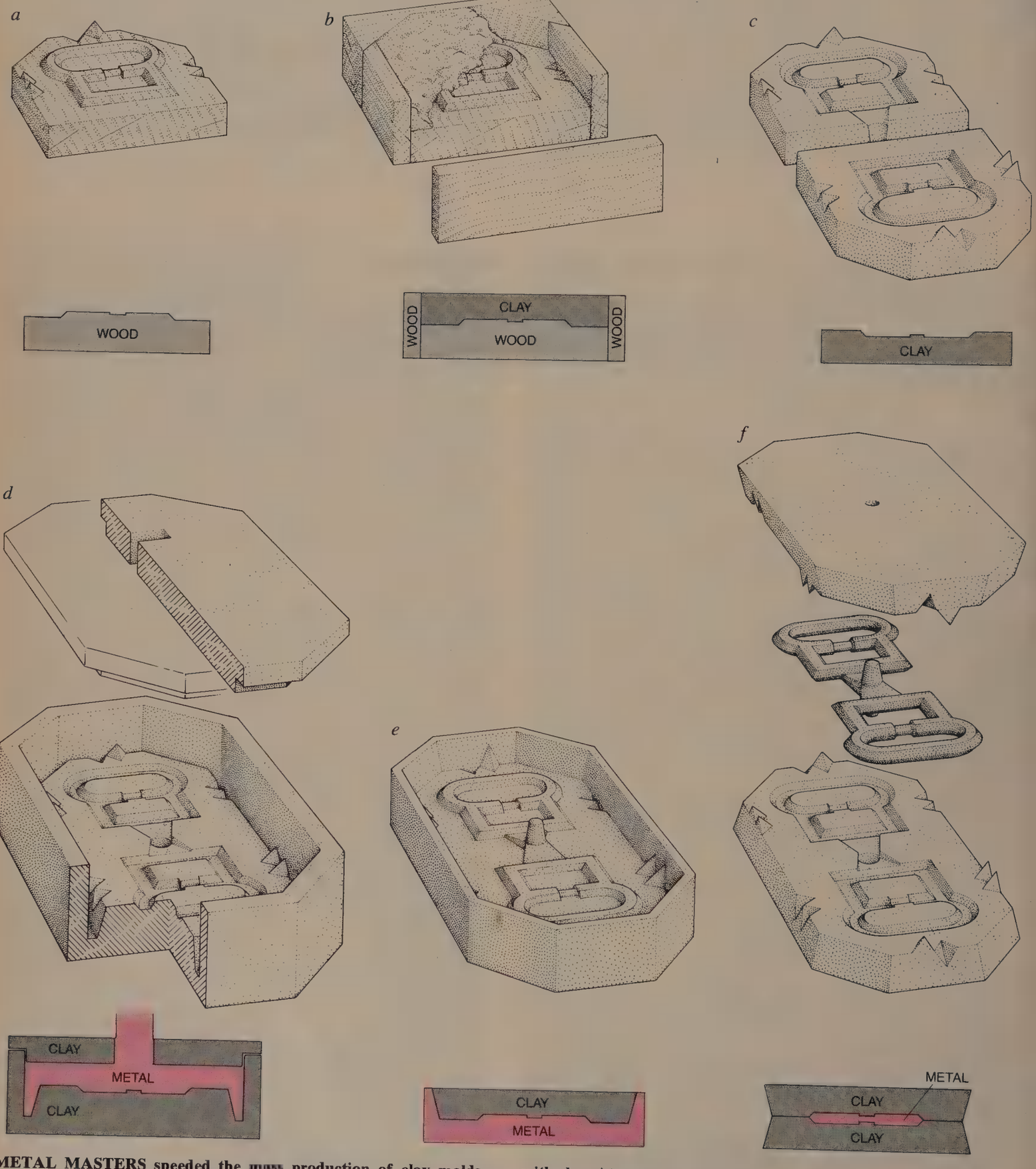
RECENT ARCHAEOLOGICAL FINDS in China include stack molds for castings (*open circles*) near Zibo in Shandong and near Nanjing and in Jurong county in Jiangsu. Molds and kilns (*dots*) have been found near Xi'an in Shaanxi and near Nanyang and Wenxian in Henan.

MATCHING HALVES of two different stack molds are shown in these photographs. The mold at the left produces two iron rings per unit. The mold at the right produces four iron buckles. Because both molds are shallow 10 or more could be stacked and filled together.

MOLDS MADE 2,000 YEARS AGO were stacked up, joined and filled by the excavators of the Wenxian kiln. The results were, at the left, 18 horse bits from a single pouring and, at the right, seven axle sleeves. Fewer of the deep axle sleeves could be poured at a time.

lated enough air-dried sets of molds to load into the kiln, they stacked the various kinds of molds together, taking care that the sprue hole in each successive mold was aligned with the hole in the one under it. Some molds were aligned by means of matching tongues and grooves; others had a hole for a guiding pin that extended through the stack. Once the molds were assembled the entire stack was coated with a layer of mud mixed with chopped straw (the standard raw material for making mud brick) and the stacks were again set aside to let the covering dry. This mud-brick covering made it possible to handle the stacked molds as a single unit. The molds were now ready for the kiln.

The Wenxian kiln proved on excavation to be three meters wide and 7.4 meters long. Along its front wall was a shal-



METAL MASTERS speeded the mass production of clay molds. The making of such a master begins at the left (*a*) with an exactly half-thick replica of a metal buckle (*top*) being carved out of wood. Next (*b*) the wood replica is reproduced in negative form by packing a box with clay. At *c* are two such negative clay reproductions. The two are then boxed together (*d*), forming the finished mold for the metal master, which is then cast. The finished metal master (*e*) is then used to make hundreds of clay molds that are paired (*f*) to cast the buckles.

Medicine is always changing, now there's a text that changes right along with it.

SCIENTIFIC AMERICAN *Medicine*.

The answers that you get from SCIENTIFIC AMERICAN *Medicine* are current, because each month new chapters replace the old. You will know where to look for the answers, because this advanced text follows the familiar structure of medical knowledge, changing as medicine changes.

Comprehensive. Leading authorities from Harvard and Stanford cover the full range of medicine for you the seasoned clinician.

Complimentary CME. At no extra cost, we offer a CME program in which you can earn 32 top credits* per year.

Trial Offer. We invite you to try SCIENTIFIC AMERICAN *Medicine* for two months with our compliments. Send us the coupon and you will receive the two-volume text and two monthly updates. You may also take a CME test for credit. At the end of 60 days, if you decide to continue the subscription, you will owe us $220 for the first 12 months (current renewal is $170); otherwise, return the books; there is no charge.

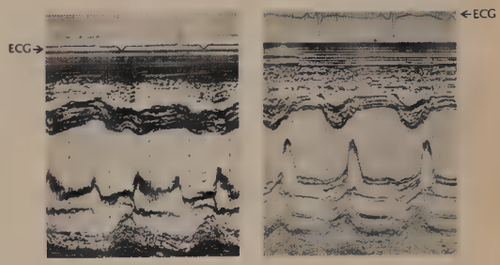
Please mail the coupon today so that we can put medical progress to work for you.

You may also order your trial copy by calling toll-free 1-800-345-8112 (in Pennsylvania, 1-800-662-2444).

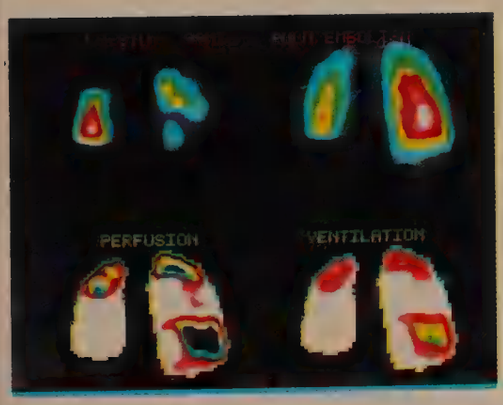
*As an organization accredited for continuing medical education, the Stanford University School of Medicine designates this continuing medical education activity as meeting the criteria for 32 credit hours in Category 1 for Educational Materials for the Physician's Recognition Award of the American Medical Association, provided it has been completed according to instructions.

This program has been reviewed and is acceptable for 32 prescribed hours by the American Academy of Family Physicians.

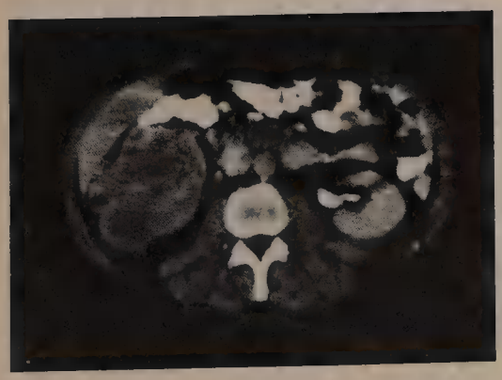
This program has been approved by the American College of Emergency Physicians for 32 hours of ACEP Category 1 credit.



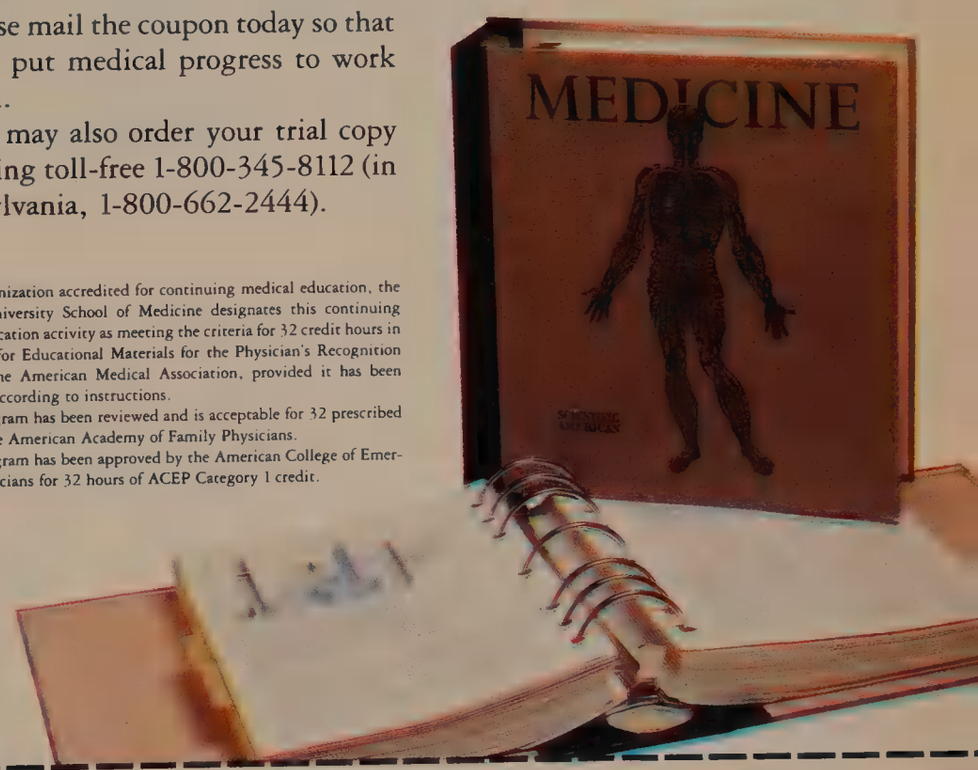
Echocardiograms from patients with aortic regurgitation.



Computerized scintigraphy reveals pulmonary thromboembolism.



Abdominal computed tomogram reveals large renal carcinoma replacing part of right kidney.



SCIENTIFIC AMERICAN MEDICINE

415 Madison Avenue, New York, New York 10017

Please enroll me as a subscriber to SCIENTIFIC AMERICAN *Medicine*. On receipt of this coupon you will send me the advanced two-volume text described in your announcement and update it regularly by sending me new monthly subsections. I understand that the annual subscription of $220 for SCIENTIFIC AMERICAN *Medicine* is tax deductible. If I am not entirely satisfied, I may cancel at any time during the first 60 days, returning all materials for a *complete refund*.

☐ Please enter my subscription for SCIENTIFIC AMERICAN *Medicine*
☐ I shall also enroll in the CME program
☐ I enclose a check made out to SCIENTIFIC AMERICAN *Medicine* for $220*
☐ Bill me ☐ VISA ☐ MasterCard

Expiration Date _____ Account Number _____

*Please add sales tax for California, Illinois, Massachusetts, Michigan and New York

Name _____ MD Specialty _____

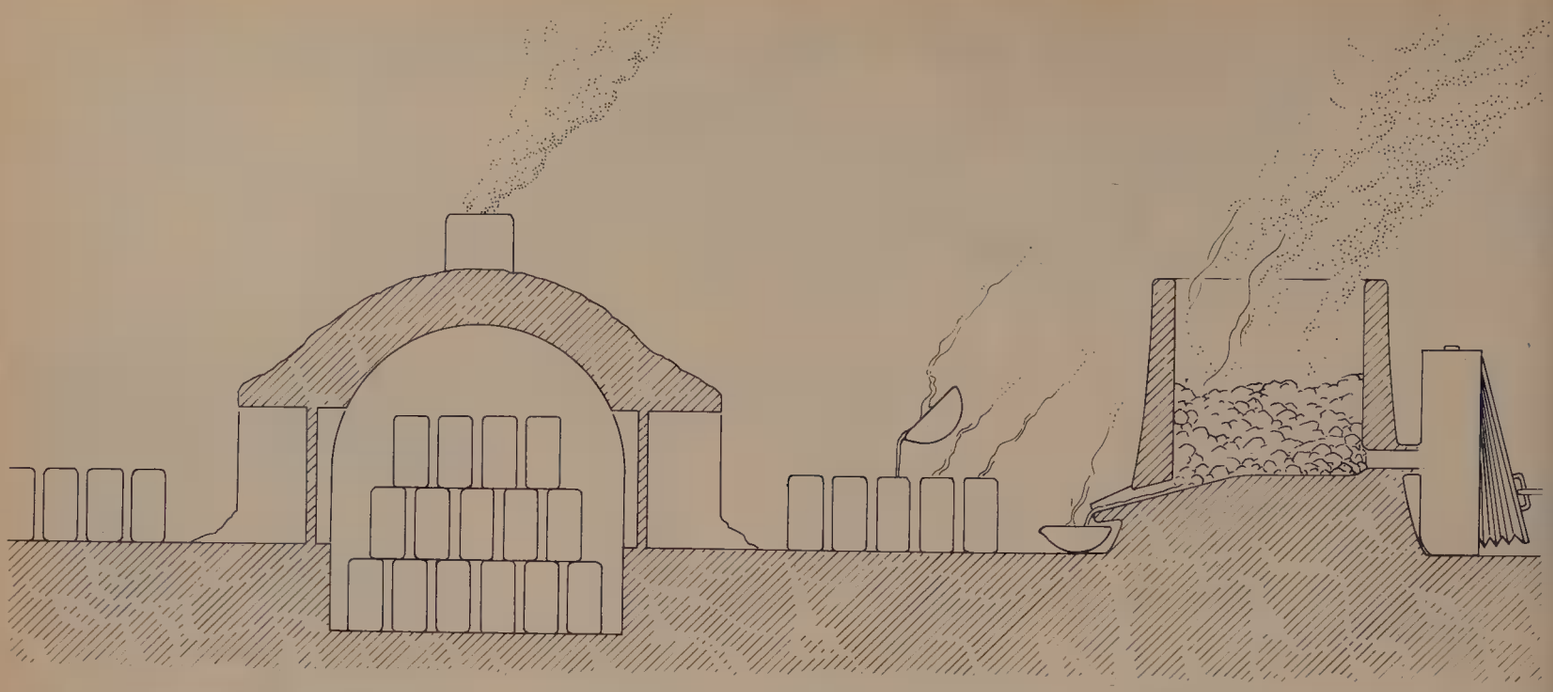
Address _____

City _____ State _____ Zip _____

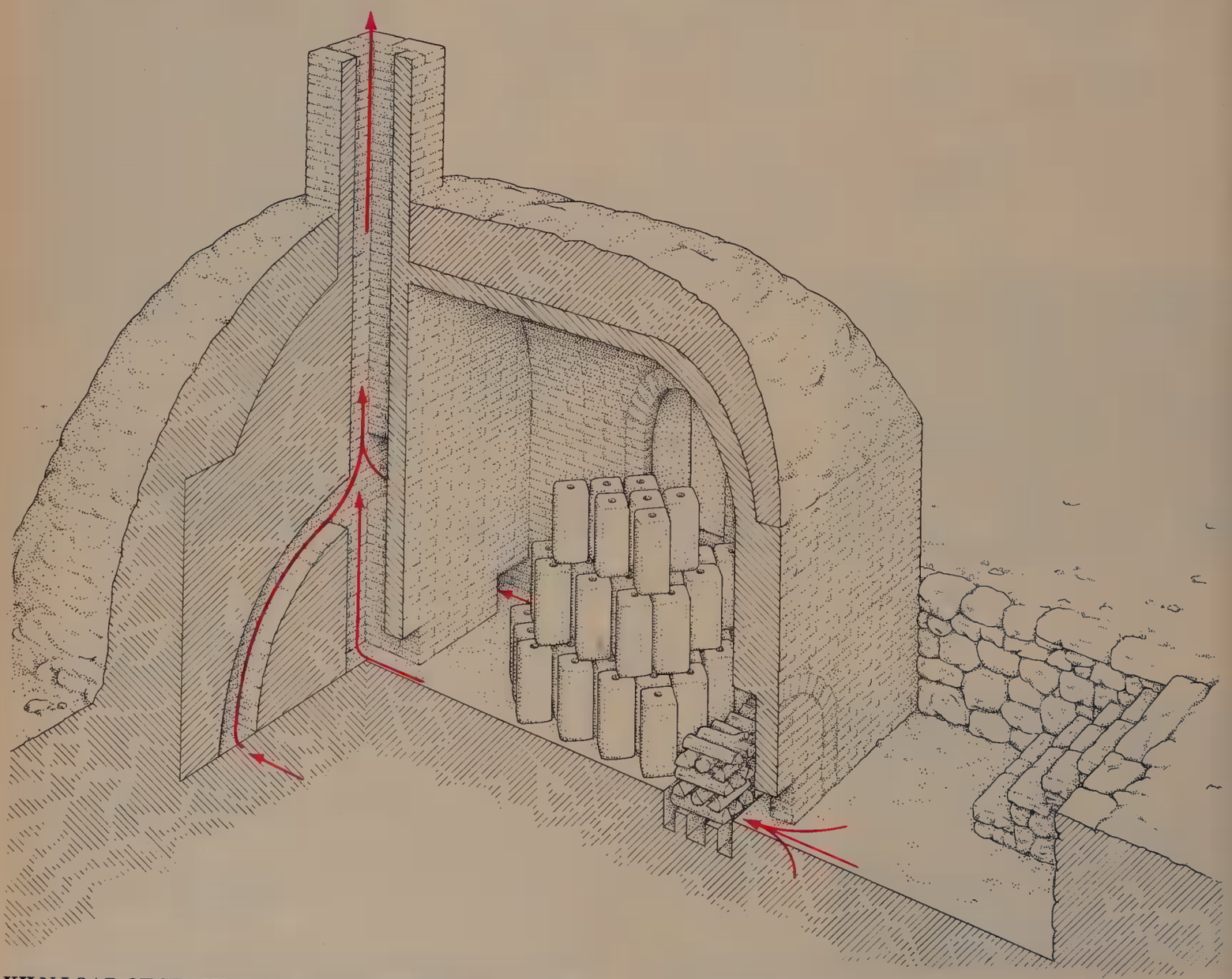
Signature _____

Please allow six to eight weeks for delivery. All payments must be in U.S. dollars. CME is available outside the U.S. and its possessions for an additional $65. Surface routes used unless airmail delivery is prepaid.

W1



POSSIBLE PRODUCTION LINE at Wenxian could have combined the kiln for heating stacked molds, at the left, with an open smelter, at the right, where the white-iron pigs could be remelted, using a bellows for forced draft. No smelter, however, was found at Wenxian.



KILN LOAD OF STACKED MOLDS is seen ready for firing in this cutaway view of the Wenxian kiln as it was restored by its excavators. The colored arrows trace the passage of hot gases from the wood fire at the right through the stacks, ensuring even heating, and out through the chimney. The kiln was three meters wide and some seven meters long. Interior temperature evidently exceeded 700 degrees Celsius.

low trench where the fire was built. The kiln floor was lined with brick; three flues in the back wall provided an outlet for the hot gases. The mud-coated mold assemblies were piled one stack on top of another and carefully positioned on the kiln floor so that the gases from the fire would flow through the spaces between the molds before reaching the exhaust flues. This evenly distributed the heat in the kiln. To judge from the color of the fired clay the temperature must have exceeded 700 degrees C.

In Han Dynasty times and for many centuries afterward the foundry fuel was wood: kindling and logs for the kilns and charcoal for the smelters. The demand eventually denuded the countryside of its forests (as it also did in Europe). As a result coal mining had become a major Chinese activity by the end of the first millennium A.D. and thereafter coal supplied most of the heating needs of industry.

THREE TIERS OF STACKED MOLDS were still standing when the excavators uncovered the site near Wenxian. A total of 500 molds were found; 16 items could be cast in 36 sizes.

Since the Han foundrymen poured the molten cast iron into their stacked molds while the molds were still hot in order to facilitate the flow of the metal, it is quite possible that a smelter or smelters operated in tandem with the mold kiln at Wenxian. In that way crucibles of liquid iron would have been available for pouring when the freshly fired stacks were ready to leave the kiln. The investigators of the site have demonstrated, however, that although this would have been economical of fuel, it was also quite practical to reheat a cooled stack some time after it was first fired: they reheated stacks nearly 2,000 years later! In the process they also learned something about the procedures probably followed by the Han foundrymen.

This information was acquired when the investigators decided to exploit some of the intact molds in order to reenact the stack-casting process. Sets of molds were cleaned and then assembled, smeared with straw mud and allowed to dry. For the first experimental casting sets of molds were kept at a temperature of 300 degrees C. for some five hours before the molten metal was poured. After cooling, the castings were found to be low in quality; the defects included large voids and excessive shrinkage. Evidently the molds had not been adequately heated. For the second experiment other sets of molds were heated to 600 degrees C. and kept at that temperature for a full six hours. By then the stacks glowed red. They were left in the oven and allowed to cool to 300 degrees C. before they were removed. Then the metal (either copper or iron) was poured into them. All the castings except one set were of high quality. The exception was a stack of key wedges. These wedges were quite thick and their molds lacked risers. Their top surface was dimpled, apparently as a result of shrinkage in the mold.

Stack casting as a mass-production technique arose in China before the middle of the first millennium B.C. The sophisticated molds found at the Wenxian site therefore represent a culmination of more than five centuries of technological progress in iron metallurgy. For example, since all the molds at the foundry were intended to be stacked for casting, their "parting" (and assembly) design was horizontal. This put a natural limit on how many objects higher than they were wide, such as the key wedges, could be cast at one pouring. Nevertheless, by stringently controlling the height of each individual mold the Wenxian foundrymen, working with a stack six molds high and with four cavities in each mold, were able to cast 24 key wedges at a time. For objects such as buckles that were longer and wider than they were high, of course, a stack could have many more individual molds and each mold could have multiple cavities. One such buckle mold, although it was not uncovered intact, carried the negative impressions of six buckles and was only a tenth as high as it was long. A stack of 20 such molds would have yielded 120 buckles per pouring.

Among the molds at Wenxian there is one of rather complex geometry that could have been intended for casting a metal master pattern. The master in turn would have yielded clay molds for casting two buckles. Making such masters was time-consuming but nonetheless efficient; it is estimated that a single master might yield as many as 10,000 clay molds before it showed signs of serious wear.

Elsewhere in China at this time simple metal molds were serving not to make clay molds but to make castings directly. Their use was confined to the casting of utilitarian objects with uncomplicated shapes: plow tips, ax and adz blades and the like. When these simple iron molds were clamped together, they formed a mold cavity complete with a pouring gate. Such "permanent" molds were thick-walled and sturdy, and their thermal conductivity was much higher than that of clay. As a result the mold could be assembled, the liquid iron could be poured, the iron could solidify, the hot casting could be removed and the mold could be reassembled for another pouring—all in a few minutes. Indeed, both the productivity and the unit cost of casting with metal molds were probably better than those of stack casting. Metal-mold casting, however, was limited to relatively simple forms. Although no such molds were found at Wenxian, the fact that castings were being made with both metal molds and clay molds in China some two millenni-

BEAUTIFUL ANTIQUE MAPS

A splendid collection of four hand-colored engravings created in facsimile from the 1660 astronomy book *Harmonia Macrocosmica* by Andreas Cellarius.

Printed in full-color using special techniques on the finest heavy stock, subjects include: The Copernican System **(shown)**; The World in 1660; Northern Constellations; Southern Constellations.

At a special introductory price of **$12.50** post paid, you'll receive all **four for the price of three!** Prints measure 19¾ by 22½, are mailed rolled and ready to frame. An informative description is included.

Add $2.00 for catalogs—$2.00 credit applied toward next purchase. Foreign orders add $3.

For credit card orders
CALL TOLL FREE 1-800-321-2369
Mon.-Fri. 8:30–4:30 Mountain Time
Utah, Alaska & Hawaii orders & inquiries call 801-535-7317

Send check, money order or VISA/MASTERCARD No. (include expiration date and signature) to:

Hansen Planetarium

Dept. SA-1 15 South State Street
Salt Lake City, Utah 84111

BINOCULARS

Precision Optics From CELESTRON®

Our Giant Series Binoculars (80mm diameter) allow you to explore the Moon and Outerspace as well as your own backyard. Celestron also offers extremely compact roof prism and classical binoculars, and superb quality Telescopes, Spotting Scopes and Telephotos. FREE product, price and ordering information by returning the coupon or send $3.00 for 52-page color catalog.

CALL TOLL FREE 1-800- 421-1526

CELESTRON INTERNATIONAL
P.O. Box 3578-SA, 2835 Columbia St.
Torrance, CA 90503

Name

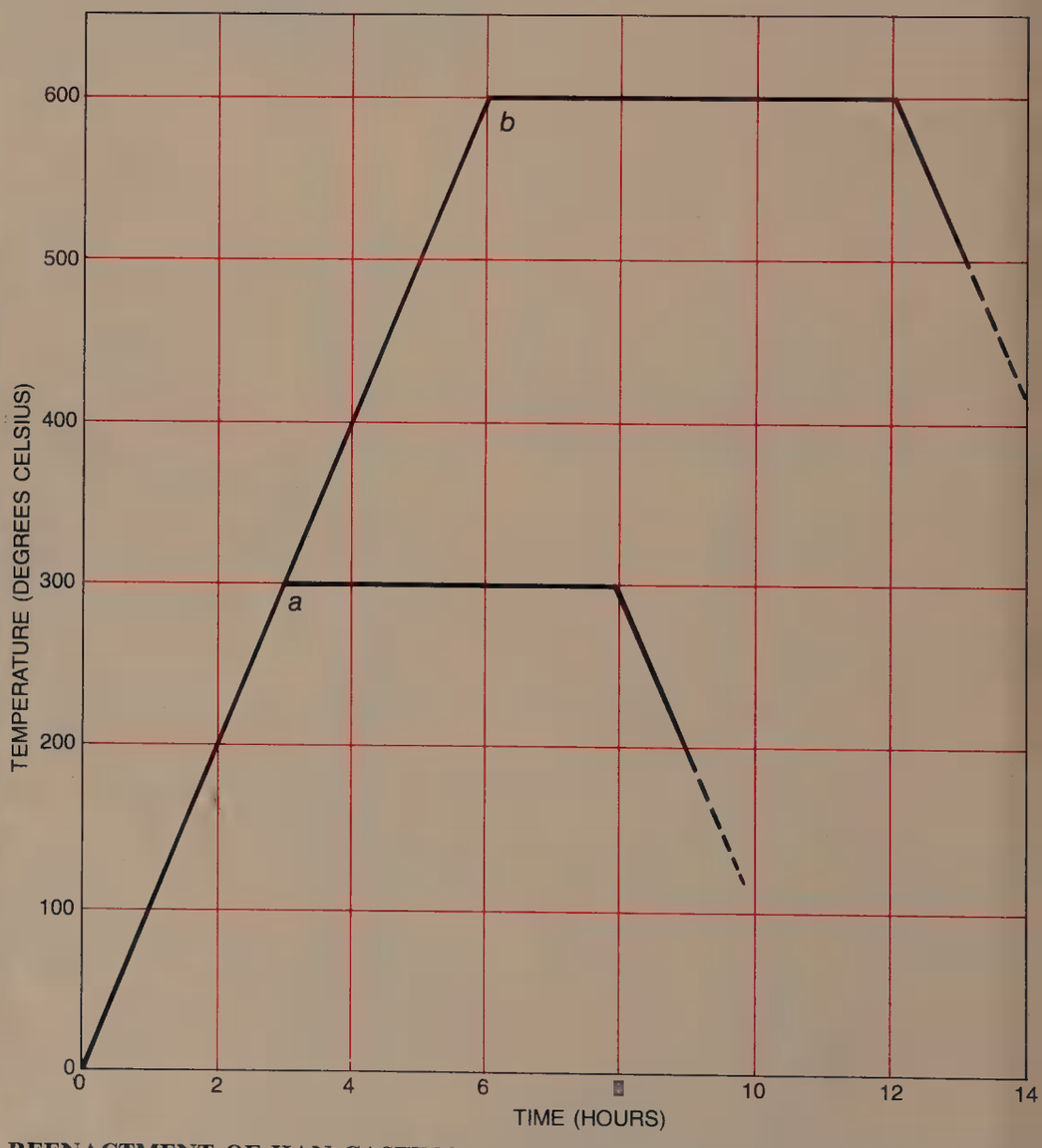
Address

City State Zip

ums ago shows that more than one good idea was being put into practice in the interests of increasing the output of the Han foundries.

What accounts for this early flowering of metallurgical ingenuity and sophistication in China? The findings of archaeologists consistently indicate that the craftsmen of the past were both intelligent and resourceful. Under the appropriate conditions any foundryman anywhere and at any time could have invented stack casting. What, then, were the conditions in China? First, the basic technologies were in place: in China both pottery manufacture and bronze casting were already ancient arts when cast-iron metallurgy began. Second, the social circumstances provided a strong stimulus for innovation: in China the technological and economic advantages of mass-producing tools, vehicle fittings and ornaments of cheap and plentiful cast iron must have been self-evident in the first millennium B.C. or there would have been no demand for them. In the light of these conditions the introduction of stack casting, although it is interesting enough in itself, perhaps more significantly suggests it would be rewarding to examine the social circumstances that nurtured such an innovation.

Stack casting not only became an important element of foundry practice early in the development of Chinese metallurgy but has continued as part of foundry practice to this day. For example, in the modern foundry at Foshan in Guangdong Province stack casting is used to produce gears and other precision metal parts notable for their accurate dimensions and fine finish. The ancient technologies uncovered at Wenxian and elsewhere not only help the Chinese of today to appreciate their heritage but also should encourage them to advance further. As the historian of technology A. B. Wilder has written: "We study the past because it is a guide to the present and a promise for the future. As we look backward our attention is directed forward."



REENACTMENT OF HAN CASTING was attempted at first (*a*) by heating the stacked molds to 300 degrees C. and heat-soaking them for somewhat less than six hours. The castings showed many defects. On a second attempt (*b*) the temperature was raised to 600 degrees and the molds were heat-soaked for a full six hours. When they had cooled to 300 degrees, the metal was poured. The quality of the castings in the second batch, with one exception, was excellent.

我们与创造明天的人接触

KE XUE is the most recently launched of SCIENTIFIC AMERICAN's six translated editions. KE XUE's paid circulation of 25,000 includes the People's Republic of China's emerging scientific, engineering and industrial leadership: the people who will make China's future happen. KE XUE is the only American originated publication on the newsstands in the PRC.

China's ability and inclination to buy will increase. They plan to accelerate progress in virtually all industrial categories: energy, aerospace, data/processing, communications, manufacturing, transportation, heavy machinery, agriculture, chemicals and mining. Now is the time to present your corporate identity, your product, your service to the People's Republic of China.

We speak the language of more than 3,000,000 world shakers in 140 countries each month. We reach the people who make the future happen in Chinese, Japanese, French, Spanish, Italian, German and English.

For more information contact:

John Kirby
Vice President
Advertising Director
SCIENTIFIC AMERICAN
415 Madison Avenue
New York, New York 10017
Telephone (212) 754-0262

SCIENTIFIC AMERICAN
54 Pall Mall
London SW 1Y 5JH, England
Telephone (011-441) 839-2873

SCIENTIFIC AMERICAN

We reach the people who make the future happen.

THE AMATEUR SCIENTIST

Simple and vivid demonstrations of advanced concepts in physics

by Jearl Walker

One might suppose that demonstrations of advanced concepts in physics would be beyond the reach of the amateur experimenter. Three such experiments have been devised, however, by Richard E. Crandall of Reed College. One experiment demonstrates the Doppler shift of light, converting the phenomenon into sound. A second measures Planck's constant, which figures importantly in quantum mechanics. The third measures the universal gravitational constant, which does the same in Newton's theory of gravitation. In each experiment Crandall has minimized the difficulties.

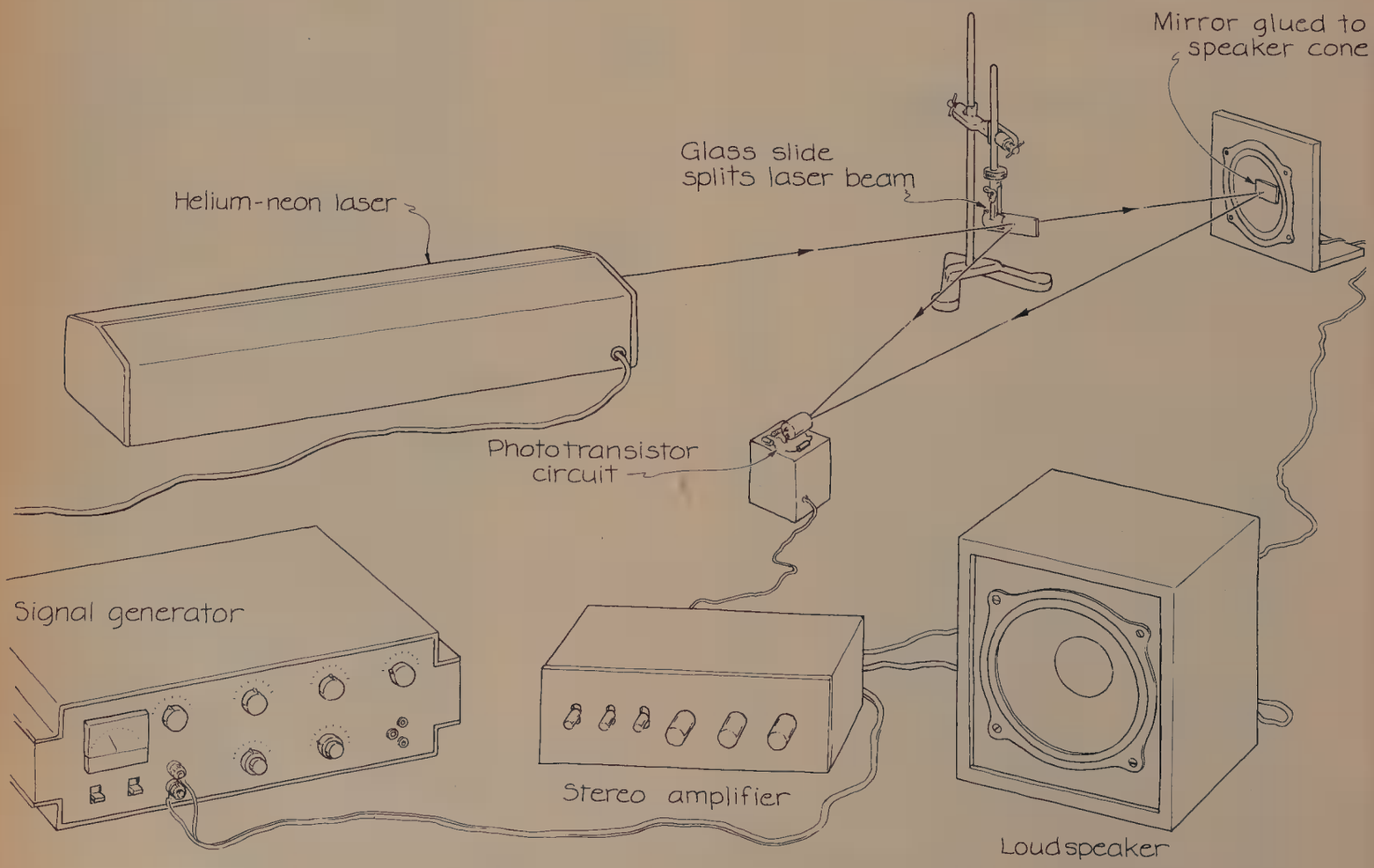
Interference patterns of light are standard fare in physics classrooms and are familiar to many amateurs. Normally the pattern is seen directly by the experimenter or is scanned with a light detector. Crandall and his student Edward H. Wishnow devised a method in which the pattern can be heard.

A beam from a two-milliwatt helium-neon laser is sent through a glass slide. Part of the beam reflects from the slide and illuminates a phototransistor (Motorola MRD3052 NPN or an equivalent type) that serves as a light detector. The rest of the beam passes through the slide and then reflects from a mirror (or anything shiny) mounted on the cone of a small loudspeaker. Part of the reflected light illuminates the phototransistor and interferes with the light arriving directly from the glass slide.

The transistor is biased (given a reference level) with a 15-volt source in line with a 30,000-ohm resistor. When the transistor is illuminated by light from the slide, it generates a voltage that is detected by a standard stereo amplifier. The amplified signal goes to a second loudspeaker that makes it audible. The other output channel of the amplifier is connected to the speaker that reflects part of the laser beam.

When the mirror moves, the signal at the transistor varies and so does the output of sound from the sound-producing speaker. Even small disturbances of the air in the room provide sufficient motion for a noticeable variation in the sound. More control of the variation is achieved by feeding a sine-wave signal into the second of the input channels of the amplifier. This signal is sent to the reflecting speaker, forcing it to oscillate steadily. The sound output from the apparatus then varies continuously.

The interference of the light at the transistor can be explained in two ways. In one view the critical factor is the difference between the distance the light travels after it is reflected from the slide and the distance it travels after it is re-



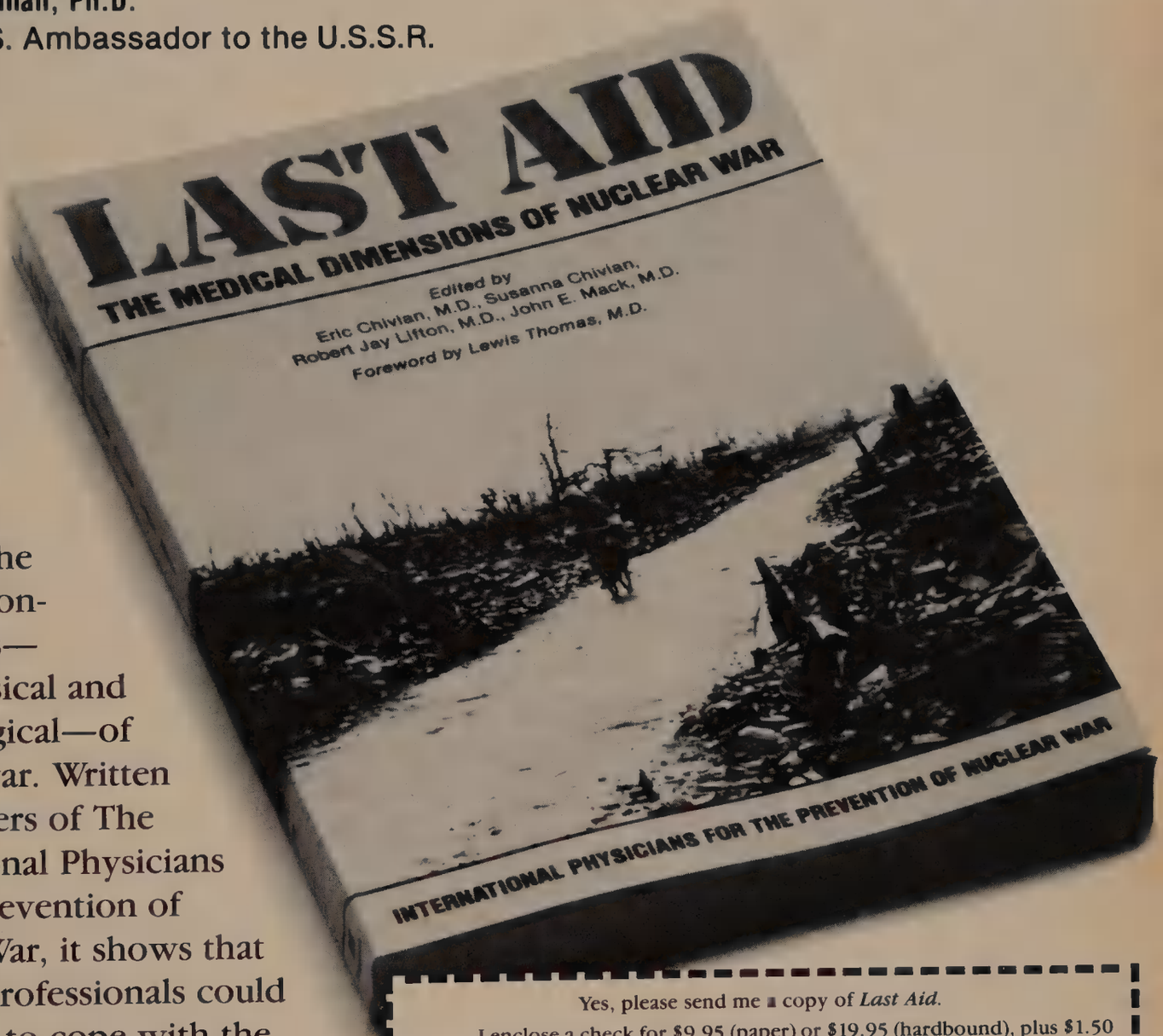
Richard E. Crandall's setup for converting the Doppler shift of light into sound

"To ignore this passionate presentation is to turn our backs on the most urgent issue of our time."

Walter Cronkite

"The facts revealed and documented in this book are fundamental for any constructive thinking about the problems of nuclear weaponry, and deserve the widest possible attention."

George F. Kennan, Ph.D.
Former U.S. Ambassador to the U.S.S.R.



Last Aid presents the medical consequences—both physical and psychological—of nuclear war. Written by members of The International Physicians for the Prevention of Nuclear War, it shows that medical professionals could not begin to cope with the overwhelming problems that would ensue in the aftermath of a nuclear holocaust. A reaction to heads of government who speak rationally of a conflict that can have no rational basis, ***Last Aid* is must reading for every intelligent, thinking human being.**

Yes, please send me ■ copy of *Last Aid*.

_____ I enclose a check for $9.95 (paper) or $19.95 (hardbound), plus $1.50 to cover postage and handling.
(California residents add appropriate sales tax.)

_____ Please charge my ☐ MasterCard ☐ VISA

Account # _____ Expires _____

Signature _____
(All credit card orders must be signed.)

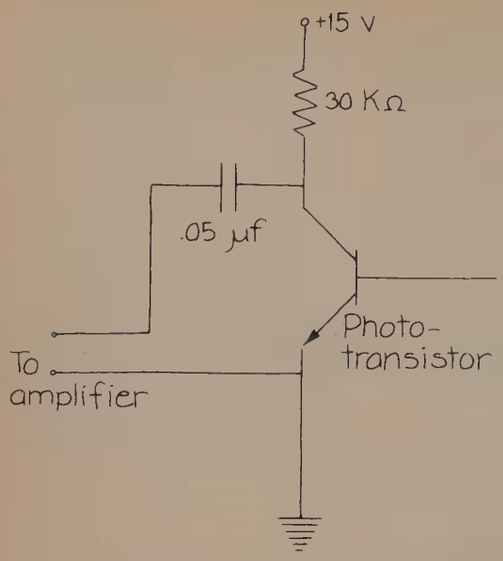
Name _____

Address _____

City _____ State _____ Zip _____

W. H. Freeman and Company
660 Market Street
San Francisco, CA 94104

ISBN hardbound 0-7167-1434-5
ISBN paper 0-7167-1435-3
SA



The detection circuit for the experiment

flected from the mirror. When the light waves emerge from the laser, they are all approximately in phase. Some of them reflect from the slide and travel to the transistor. The rest of the waves travel a different path to reach the mirror and the transistor.

The difference between the length of the path traveled by one set of waves and the length of the path traveled by the other determines the interference at the transistor. If the path lengths are such that the two sets of waves arrive in phase, the interference is constructive: the illumination of the transistor is bright and the transistor puts out a strong signal to the amplifier. If the path lengths are such that the two sets arrive out of phase, the interference is destructive: the illumination of the transistor is dark and a weaker signal is sent to the amplifier.

Suppose the first light waves arrive in phase. Also suppose the mirror moves away from the laser by a distance equal to one-fourth of the wavelength of the laser light. The light reflecting from the speaker then has an additional distance to travel that is equal to half of the wavelength of the light. (It must travel through an extra one-fourth wavelength to reach the speaker and then back again to reach the transistor.) The additional distance puts the light waves at the transistor out of phase. Thus as the cone of the reflecting speaker moves parallel to the light beam it varies the strength of the signal from the transistor.

One can hear the random motion of the mirror as it is disturbed by slight movements of air in the room. When a sine-wave signal drives it in a more controlled oscillation, the variation in the interference at the transistor generates a repetitive pattern. The mirror moves through much more than one-fourth of the wavelength of the light, so that during each oscillation the interference at the transistor reaches many strong and weak levels.

A different but related explanation of the experiment is based on the Doppler shift of light. Wavelength is related to frequency: the number of waves that pass per second. The frequency of the light perceived by an observer can depend partly on his motion with respect to the source of the light. Assume the observer is stationary with respect to the light source. He then perceives light at a certain frequency. If the source moves toward the observer, the perceived light is shifted to a higher frequency. This shift is often called a blue shift because it is toward a higher frequency and so toward the blue end of the spectrum of visible light. If the source moves away from the observer, the perceived light is shifted to a lower frequency. This shift is often called a red shift. Both blue and red shifts are Doppler shifts.

A Doppler shift in frequency also results, of course, if the observer moves toward or away from the source. The degree of the shift depends on the relative speed of the source and the observer. The closer that speed is to the speed of light, the larger the Doppler shift is. This fact is crucial in inferring the motion of a distant astronomical object. The object's Doppler-shifted light can be measured and its speed toward or away from the earth can be calculated.

In the experiment done by Crandall and Wishnow the mirror moves relative to the transistor at a speed that is low compared with the speed of light. The resulting shift in the frequency of the light is quite small. The light is actually shifted twice in frequency, once because the mirror moves relative to the laser and again because it moves relative to the transistor. In the first instance the light source is the laser and the mirror is the observer. In the second the mirror acts as a source (since it is sending out light) and the transistor is the observer.

Normally the special theory of relativity is required to calculate the Doppler shift of light. The speed of the oscillatory motions of the loudspeaker is so low compared with the speed of light, however, that an approximation can be made on the basis of classical physics. The shift in frequency is equal to twice the unshifted frequency multiplied by the ratio of the speed of the mirror to the speed of light.

The light arriving at the transistor from the glass slide is unshifted. The light from the moving mirror is shifted up or down in frequency depending on which way the mirror is moving when the light reaches it. These two signals interfere with each other at the transistor to create what is called a beat frequency. The net illumination at the transistor varies between constructive and destructive interference with a beat frequency that is equal to the difference between the unshifted frequency of the light and the shifted frequency. The experimenter hears this beating as a variation in the sound emerging from the sound-producing speaker. In other words, he hears a pattern that is related to the Doppler shift of light.

Stefan's law ——— $P = \epsilon A \sigma T^4$

Stefan's constant ——— $\sigma = \dfrac{2\pi^5 k^4}{15 h^3 c^2}$

Planck's equation ——— $I = \dfrac{Nf^3}{\exp(hf/kT) - 1}$

Approximate equation for Planck's constant ——— $h \simeq \dfrac{15c^2}{2\pi^5 A f^4} \cdot \dfrac{P_1 P_2}{(P_1^{1/4} - P_2^{1/4})^4} \cdot \mathrm{Log}^4 \left(\dfrac{I_1}{I_2}\right)$

Equations for determining Planck's constant

At the end of the 19th century the subject of thermal radiation presented a serious problem in physics. Every body radiates electromagnetic waves according to the temperature of its surface. Some of the radiation might be in the visible range of the electromagnetic spectrum, so that it can be seen, but every surface also emits radiation over a broader range of the spectrum. For example, a poker heated in a fireplace gives off infrared radiation that can be felt long before the surface is hot enough to glow in the visible range.

Armed with classical thermodynamics and the fresh brilliance of James Clerk Maxwell's theory of electromagnetic radiation, physicists attempted to derive an expression for the intensity of the thermal radiation at a particular frequency. Repeated efforts yielded only approximations. At low frequencies a mathematical expression derived by Lord Rayleigh approximated the ex-

THE INPUT/OUTPUT STRUCTURE OF THE UNITED STATES ECONOMY

WHAT MAKES THE U.S. ECONOMY TICK?

The editors of SCIENTIFIC AMERICAN have prepared a wall chart displaying for the 1980's the Input/Output Structure of the U.S. Economy based on the latest interindustry study from the U.S. Department of Commerce.

The SCIENTIFIC AMERICAN Input/Output wall chart does for economics what the table of elements does for chemistry. It answers at a glance questions about the linkage between the microeconomics of the firm and the macroeconomics of the system; about the web of technological interdependencies that tie industry to industry; about the industry-by-industry direct and indirect consequences of swings in public and private spending; about the impact of change in technology, and about any other topic you can think of. You are rewarded by surprise as well as by confirmation of your hunches. For teaching and practical and theoretical studies, here is a powerful, graphic tool.

In the familiar format of the SCIENTIFIC AMERICAN Input/Output wall charts for the 1960's and 1970's, the wall chart for the 1980's measures 65″ × 52″ and is printed in eight colors. Each of the nearly 10,000 cells in the 97-sector interindustry matrix shows (1) the interindustry commodity flow, (2) the direct input/output coefficient and (3) the "inverse" coefficient. Where the direct input/output coefficient exceeds .01, the cell is tinted in the color code of the industrial bloc from which the input comes. This device, combined with triangulation of the matrix, brings the structure of interindustry transactions into graphic visibility.

A supplementary table displays, industry by industry, the capital stock employed; the employment of managerial, technical-professional, white-collar and blue-collar personnel; the energy consumption by major categories of fuel, and environmental stress measured by tons of pollutants.

The editors of SCIENTIFIC AMERICAN are happy to acknowledge the collaboration, in the preparation of this wall chart, of Wassily Leontief, originator of input/output analysis—for which contribution to the intellectual apparatus of economics he received the 1973 Nobel prize—and director of the Institute for Economic Analysis at New York University.

Packaged with the chart is an index showing the BEA and SIC code industries aggregated in each of the 97 sectors.

SCIENTIFIC AMERICAN

415 Madison Avenue, Dept. I/O, New York, N.Y. 10017

YES, send me the new SCIENTIFIC AMERICAN 1980's INPUT/OUTPUT wall chart—printed in eight colors on heavy paper measuring 65″ × 52″.

I understand my chart(s) will be shipped prepaid, uncreased*, flat-rolled and in a sturdy mailing tube. (Allow 4-6 weeks for delivery)

I enclose a check for $_____ for _____ chart(s) at $35 each, saving me $5 each in handling charges.

Bill me for $_____ for _____ chart(s) at $40 each.

Name_____

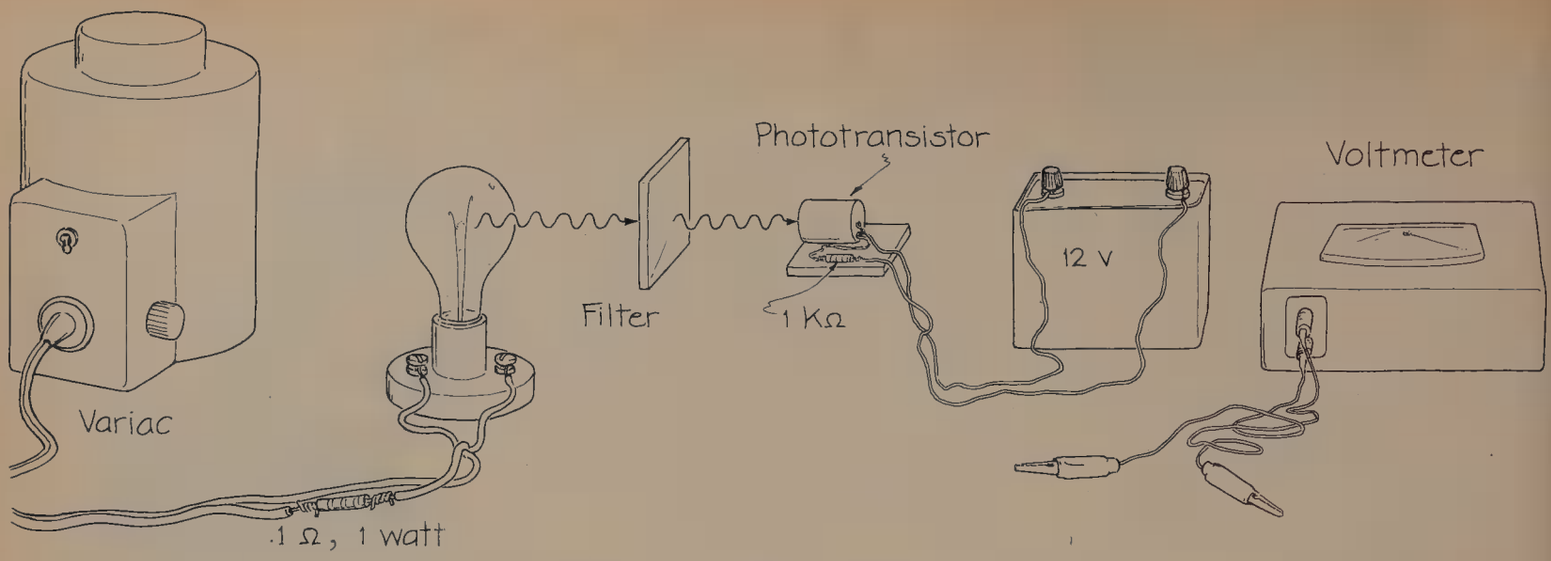
Company_____

Street Address_____

City, State, Zip Code_____

*Charts shipped international require one fold. 1/83

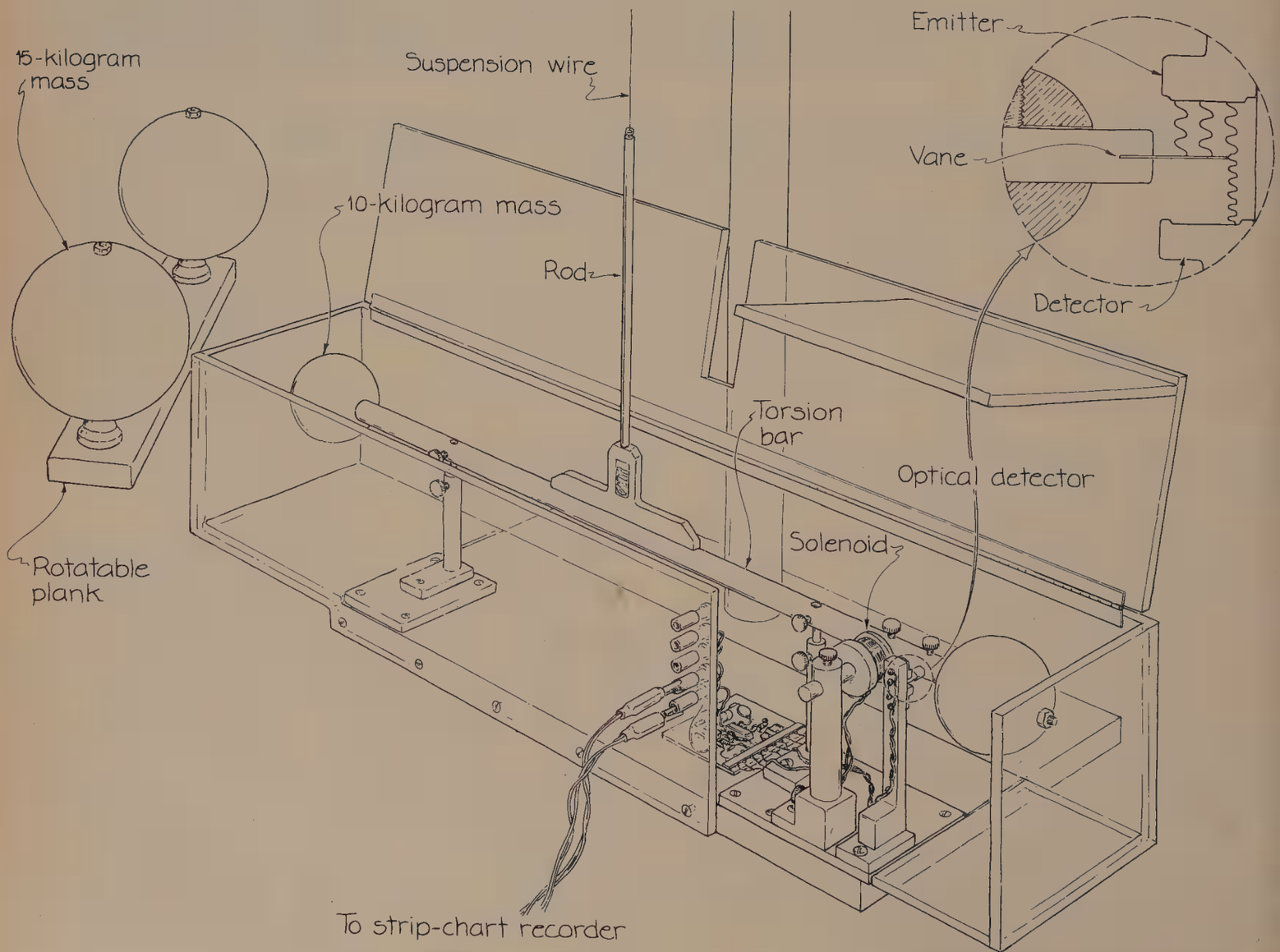
Make check payable to SCIENTIFIC AMERICAN. Please add sales tax for California, Illinois, Massachusetts, Michigan and New York.



How Crandall measures Planck's constant

perimental results, but at high frequencies the expression failed. Indeed, it predicted that the power radiated at high frequencies would be infinitely large, which was clearly wrong. This mathematical disaster was known as the ultraviolet catastrophe. Earlier Wilhelm Wien had devised an expression that worked well at high frequencies but was obviously wrong at low frequencies.

The problem was resolved in 1900 when Max Planck presented an equation that closely agreed with the experimental evidence at all frequencies. The equation was almost a product of luck, because it was not until several weeks later that Planck understood why it worked. In his theoretical model of a radiating surface he imagined that the radiation was emitted from many small oscillators. The intensity of the radiation at a given frequency was simply the combined output from a distribution of these oscillators. The surprising feature



Crandall's arrangement for measuring the universal constant of gravity

of the model was that the oscillators could not have just any energy on a scale of energies. Instead each oscillator was limited to particular energy values that were integral multiples of a fundamental value given by the product of a constant (now called Planck's constant h) and the frequency f of the oscillator.

Planck had no information about the atomic nature of a radiating surface, and his imaginary oscillators certainly did not exist. With his crucial discovery, however, he was able to justify his equation for the thermal radiation as a function of frequency. It was more important that Planck's work ushered in the age of quantum mechanics. On the microscopic scale energy values were limited to certain distinct quantities that were integral multiples of hf.

Although Planck's constant is an important value in the most advanced quantum mechanics of atoms and fundamental particles, it can be measured in a surprisingly simple experiment devised by Crandall and his colleague Jean F. Delord. The apparatus consists of a color filter, a light detector and an unfrosted 60-watt tungsten-filament light bulb. Light from the filament of the bulb is filtered so that only a narrow range of frequencies reaches the detector. The intensity of the light in that range is measured for two levels of power delivered to the light bulb, and the results are inserted in an equation that yields a value for Planck's constant.

The filament of the light bulb is heated by the passage of current and emits light across a broad band of frequencies in the visible and infrared ranges. The total power radiated from a surface was computed by Josef Stefan in 1879 as being equal to the product of four factors: (1) the surface area, (2) a constant known as Stefan's constant, (3) the absolute temperature of the surface raised to the fourth power and (4) a value known as the emissivity of the material.

At given values of surface area and temperature a material can theoretically radiate a certain maximum amount of power. Such a material is assigned an emissivity of 1 and is called a black-body radiator. No such ideal material exists; all real surfaces radiate less power. Each surface is assigned an emissivity smaller than 1 according to how well it radiates. The filament in the light bulb employed in the experiment runs hot enough to be almost a black-body radiator. Hence Crandall and Delord have made the approximation of assigning it an emissivity of 1.

Planck's constant appears in the modern expression of Stefan's constant, so that it is related to the total power radiated by the filament. It also appears in the equation devised by Planck for the intensity of the light radiated at any given frequency. Crandall and Delord perform two sets of measurements in order to get a value for Planck's constant. In each set they find the total power radiated by the filament by measuring the electric power delivered to the bulb. They also measure the intensity of the light passing through the color filter. This set of measurements is then repeated for a different amount of power delivered to the bulb. The data are inserted into an equation that is the combination of Stefan's law (the total power radiated) and the equation from Planck (the intensity of the light at a certain frequency).

The combined equation is shown in the bottom illustration on page 132. The equation incorporates an approximation in order to simplify the mathematics, but the accuracy of the final value of Planck's constant is hardly altered by it. The symbols I and P represent the intensity and power Crandall and Delord measure in the experiment. The equation also requires a value for the speed of light c, which is 3×10^8 meters per second. The symbol f is the frequency of the light passed by the filter. Since the filter actually passes not one frequency but a narrow range of frequencies, the value for this factor is chosen as being the center of the range. If the range were narrower, less error would be introduced by that approximation.

The symbol A in the equation is the surface area of the filament. To measure it Crandall and Delord project a shadow image of the filament onto a wall. The dimensions of the image are measured and then scaled down by the magnification of the projection. You can determine the magnification easily by placing a transparent ruler near the bulb and measuring its image on the wall. The value for A in their experiment was 5.2×10^{-5} square meter.

A Variac controls the power delivered to the bulb. To calculate the power Crandall and Delord measure the voltage across the .1-ohm resistor that is in line with the bulb. Dividing the voltage by the resistance gives the current flowing through the resistor and thus also through the filament.

Next Crandall and Delord measure the voltage across the bulb. Multiplying this voltage by the current gives them the power delivered to the bulb. Crandall says that if a Variac is not available, a 12-volt battery can be substituted, but one would then have to adjust the current with an appropriate resistor in order to deliver two different levels of power to the bulb.

The filter placed between the bulb and the photodetector is a green cellophane one bought from Corion Instruments, Inc. (73 Jeffrey Avenue, Holliston, Mass. 01746). The frequency entered in the calculation of Planck's constant was 5.3×10^{14} hertz. Similar filters can be obtained from other scientific-supply companies. Different colors can be substituted. The band pass, that is, the frequency range, of the filter should be as narrow as possible, but even an inexpensive filter will serve for determining a good value for Planck's constant.

The detection circuit consists of a phototransistor (Motorola MRD3052 or its equivalent), a 1,000-ohm resistor and a 12-volt battery. The spectral response of the phototransistor is not important because the frequency of the light illuminating the transistor does not vary. The current in the detection circuit is determined by measuring the voltage drop across the resistor. The current need not be converted into units of light intensity. Since a ratio of only two intensities is required in the equation, the units cancel.

Crandall sent me the following example of the data. With a power input of 19.3 watts to the bulb the current in the detection circuit was .81 milliampere. When the power input was changed to 36.5 watts, the current changed to 7.57 milliamperes. From these data Crandall and Delord calculated Planck's constant to be 4.9×10^{-34} joule-second, which is close to the accepted value of 6.6×10^{-34} joule-second. Their value is off somewhat because of the finite band pass of the filter and because the emissivity of the filament is not exactly 1.

Of the several fundamental forces in the universe the most noticeable in everyday life is gravitation. The gravitational pull on one's body from the presence of the earth is appreciable only because of the large mass of the earth. The gravitational attraction between any pair of common objects around the house is so small that it goes unnoticed. It is overwhelmed by much larger forces such as friction.

How does one demonstrate and measure the gravitational force between two masses in a laboratory? This question was first tackled by Henry Cavendish in 1798. His apparatus consisted of a six-foot wood arm suspended horizontally by a thin wire. At each end of the arm he mounted a lead ball about two inches in diameter. The suspension was extremely sensitive to any force tending to rotate the arm about the suspension wire.

To present such a force a third lead ball was brought near one of the mounted balls. Its gravitational pull on the mounted ball exerted a force tending to rotate the arm about the suspension wire. To increase the force a fourth lead ball was brought on the other side of the ball at the other end of the arm.

Cavendish was careful to shield the apparatus from air currents. He also arranged for the third and fourth balls to be moved into position by remote control so that the mass of his own body

would not interfere with the experiment. The angle of rotation was measured by monitoring a light beam reflected from a small mirror on the wire from which the beam was suspended.

With the third and fourth lead balls in position the arm rotated slightly until the gravitational pull was countered by the torque from the twisted wire. This position of equilibrium was measured. Then the third and fourth balls were repositioned so that they exerted forces in the opposite direction. Again the angle of equilibrium was measured. The angular difference between the two positions of equilibrium was divided by two to give an accurate measure of the rotation in each case.

With these numbers Cavendish was able to compute the strength of the gravitational force between the pairs of lead balls at each end of the arm. From work by Isaac Newton the mathematical form of the gravitational force between two masses was already known. Its strength depends directly on the two masses attracting each other and inversely on the square of the distance between them. It also depends on the factor called the gravitational constant. Newton was effectively stymied on how to compute the constant because his only demonstration of gravitation was based on at least one large object whose mass was not well known (such as the earth). The Cavendish experiment was exciting because it was based on smaller masses that could be measured easily. Cavendish computed the gravitational constant as being $6.754 \times 10^{-11} N(m^2/\text{kg.}^2)$, close to the currently accepted value of $(6.668 \pm .005) \times 10^{-11} N(m^2/\text{kg.}^2)$. ($N$ is newtons, m is meters, kg. is kilograms.)

The Cavendish experiment is not an ordinary demonstration in physics because it is arduous to do. The arm takes a good deal of time to settle down into its equilibrium position. Any perturbation from the environment (for example from the experimenters' moving around the laboratory) adds further to the settling time. As a result the demonstration often takes hours. Working from a prototype built by Greg Eibel of Reed College, Crandall has devised a clever electronic version of the Cavendish experiment that takes only minutes and yields a value for the gravitational constant of $(7.5 \pm 1.5) \times 10^{-11} N(m^2/\text{kg.}^2)$.

Whereas the classical Cavendish experiment depends on the rotation of an arm, the experiment devised by Crandall depends on preventing such a rotation. The strength of the gravitational force is measured from the magnetic force that is required to prevent the rotation of a torsion bar when an additional mass is brought close to it. The position of the torsion bar is monitored by an optical sensor.

When the additional mass is brought near one of the masses mounted on the bar, the start of rotation by the bar triggers a servomechanism that exerts enough magnetic force on the bar to stop the rotation. By calibrating the servomechanism Crandall gets a measure of the gravitational pull on the mass mounted on the bar. The entire experiment can take as little as 20 minutes.

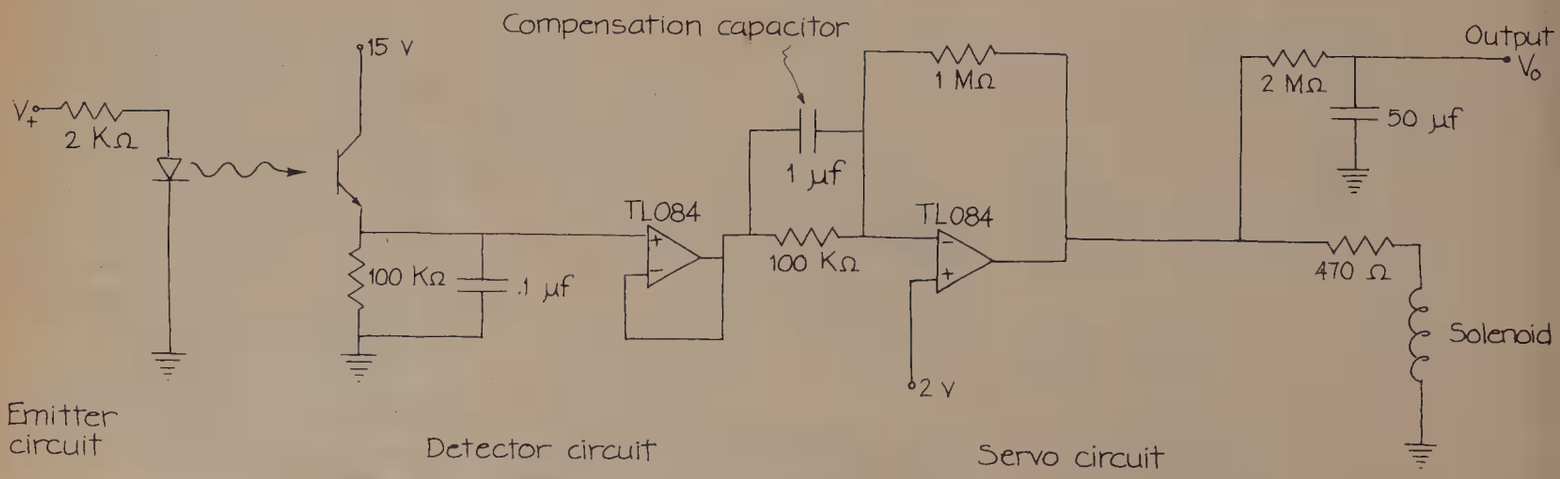
The bar is suspended by a length of piano wire that runs from an overhead mount down to a rigid rod attached to the middle of the bar. The rod is not necessary, but it helps to damp out vibrations of the wire. Masses of 10 kilograms are mounted at both ends of the bar as in the classical apparatus. Near one end a magnet is glued into a hole extending through the bar. When the bar begins to rotate, a solenoid in the servomechanism pulls magnetically on the magnet to eliminate the rotation.

The optical sensor consists of an infrared emitter and an infrared detector positioned near the end of the bar with the magnet. The circuitry of the apparatus is shown in the illustration below. The emitter is a simple diode that radiates a small amount of heat as current from a constant-voltage source flows through it. The detector is a transistor with electrical characteristics that change when it is exposed to heat. (The combined emitter and detector can be bought as an optical interrupter unit, one such being the General Electric H13A1.)

Between the emitter and the detector a vane is attached to the bar. Initially the vane shields half of the detector from the radiation. As the bar begins to rotate in one direction the detector is more exposed and the servomechanism acts to restore the bar to its initial orientation. Rotation in the other direction shields more of the detector and again brings the servomechanism into play.

The output of the detector passes through a network of two operational amplifiers and other devices. The system constantly compares the detector's output to a reference voltage of two volts. If the bar is in its proper position and the vane shields half of the detector from the infrared emitter, the output matches the reference voltage. If the bar rotates so as to expose the infrared detector, the match is disrupted and the system sends current through the solenoid, generating a magnetic field that pulls on the magnet attached to the bar. The amount of current required for the correction is read indirectly from the change in the output voltage (V_o in the illustration). This output appears on a voltmeter or a strip-chart recorder.

The first step in calibrating the apparatus is to calculate the torsion constant of the wire from which the bar is suspended. When the bar is rotated through an angle, the twisted wire supplies a countering torque equal to its torsion constant multiplied by the angle of rotation. To compute the torsion constant Crandall determines two numbers. With the servomechanism turned off he makes the bar swing around its equilibrium position and measures the period of swinging. This amount of time is called the free period of the bar. Next the bar's moment of inertia (with respect to the point of suspension) is computed. The torsion constant of the wire is then calculated by the formula $k = I(2\pi/P)^2$, where k is the torsion constant, I is



The circuitry of the experiment with gravity

the moment of inertia and P is the free period.

The next step is to relate the output voltage from the servomechanism to a torque on the bar. With the servomechanism turned on the support for the suspension wire is moved through a measured angle in order to twist the wire. The bar remains in place because of the magnetic pull from the solenoid. Thus the output voltage from the circuit shifts. The voltage shift is divided by the angle of twist and the torsion constant to yield what Crandall calls the torsional gain of the apparatus. This factor (volts per unit of torque) allows him to convert any output voltage measured during an experiment into the value of the torque on the bar because of the gravitational pull on one end of the bar.

For example, when a 15-kilogram mass almost touches one of the mounted masses, the output voltage changes by about 25 millivolts. The strength of the force on the mounted mass is computed by multiplying the shift in the output voltage by half the length of the bar and dividing by the torsional gain. The gravitational constant is calculated by inserting this force into Newton's equation for gravitation, along with the distance between the center of the 15-kilogram mass and the center of the mounted mass.

To ease the introduction of an additional mass Crandall sometimes employs a rotatable plank with a 15-kilogram mass mounted at each end. It functions something like the revolving tray known as a lazy Susan. By rotating the plank in one direction one of the masses is brought to a side of the mass mounted on the bar. The bar is pulled toward the additional mass but barely moves because of the servomechanism. When the plank is rotated the other way, the other 15-kilogram mass is brought to the opposite side of the mounted mass.

The settling time for the apparatus is set by the compensation capacitor, which has a value of about one microfarad. When Crandall brings a 15-kilogram mass near one end of the bar, the shift in the output voltage appears after only a few minutes. This speed makes Crandall's version of the Cavendish experiment ideal for a classroom demonstration. An accurate measurement of the gravitational constant calls for only 20 minutes or so of observation.

For his work on the gravitational constant Crandall won the 1981 Apparatus Competition of the American Association of Physics Teachers. This experiment and the one measuring Planck's constant will appear soon in *American Journal of Physics*. The experiment with the Doppler shift of light has already been published there. Crandall's colleague Bruce Eaton contributed ideas to some of these projects.

Now NRI takes you inside the new TRS-80 Model III microcomputer to train you at home as the new breed of computer specialist!

NRI teams up with Radio Shack advanced technology to teach you how to use, program and service state-of-the-art microcomputers...

It's no longer enough to be just a programmer or a technician. With microcomputers moving into the fabric of our lives (over 250,000 of the TRS-80™ alone have been sold), interdisciplinary skills are demanded. And NRI can prepare you with the first course of its kind, covering the complete world of the microcomputer.

Learn At Home In Your Spare Time

With NRI training, the programmer gains practical knowledge of hardware, enabling him to design simpler, more effective programs. And, with advanced programming skills, the technician can test and debug systems quickly and easily.

Only NRI gives you both kinds of training with the convenience of home study. No classroom pressures, no night school, no gasoline wasted. You learn at your convenience, at your own pace. Yet you're always backed by the NRI staff and your instructor, answering questions and giving you guidance.

You Get Your Own Computer to Learn On and Keep

NRI training is hands-on training with practical experiments and demonstrations. You don't just program your computer, you go inside it...watch how circuits interact...interface with other systems ...gain a real insight into its nature.

You also work with an advanced liquid crystal display hand-held multimeter and the NRI Discovery Lab,® performing over 60 separate experiments. Both microcomputer and equipment come as part of your training for you to use and keep.

Computer Assisted Instruction

Your TRS-80 even helps train you. You receive 8 special lesson tapes in BASIC computer language. Using them in your microcomputer, you "talk" to it as you progress. Errors are explained, graphics and animation drive home key points. Within a matter of minutes, you'll be able to write simple programs yourself.

Send for Free Catalog... No Salesman Will Call

Get all the details on this exciting course in NRI's free, 100-page catalog. It shows all equipment, lesson outlines, and facts on other electronics courses such as Electronic Design, Industrial Electronics, TV/Audio/Video Servicing...11 different career opportunities in all. Send today, no salesman will ever bother you. Keep up with the latest technology as you learn on the latest model of the world's most popular computer. If coupon has been used, write to NRI Schools, 3939 Wisconsin Ave., Washington, D.C. 20016.

(TRS-80 is a trademark of the Radio Shack division of Tandy Corp.)

NRI McGraw-Hill

NRI Schools
McGraw-Hill Continuing Education Center
3939 Wisconsin Avenue
Washington, D.C. 20016
We'll give you tomorrow.

NO SALESMAN WILL CALL.

Name (Please Print) Age

Street

City/State/Zip

Accredited by the Accrediting Commission of the National Home Study Council

Please check for one free catalog only.
- ☐ Computer Electronics including Microcomputers
- ☐ Color TV, Audio, and Video System Servicing
- ☐ Electronics Design Technology
- ☐ Digital Electronics
- ☐ Communications Electronics • FCC Licenses • Mobile CB • Aircraft • Marine
- ☐ Industrial Electronics
- ☐ Basic Electronics
- ☐ Small Engine Servicing
- ☐ Appliance Servicing
- ☐ Automotive Servicing
- ☐ Auto Air Conditioning
- ☐ Air Conditioning, Heating, Refrigeration, & Solar Technology
- ☐ Building Construction

All career courses approved under GI bill.
☐ Check for details

183-013

BIBLIOGRAPHY

Readers interested in further explanation of the subjects covered by the articles in this issue may find the following lists of publications helpful.

METAMAGICAL THEMAS

MIND, BRAIN, AND HUMANIST VALUES. Roger Sperry in *New Views on the Nature of Man,* edited by John R. Platt. University of Chicago Press, 1965.

THE SELFISH GENE. Richard Dawkins. Oxford University Press, 1976.

THE SCHEME OF THINGS. Allen Wheelis. Harcourt Brace Jovanovich, Inc., 1980.

CHANCE AND NECESSITY: AN ESSAY ON THE NATURAL PHILOSOPHY OF MODERN BIOLOGY. Jacques Monod. Alfred A. Knopf, Inc., 1981.

THETAMAGICAL MEMAS: SENTENCE-LIKE VIRUSES AND SELF-STRUCTURING REPLICATORS. Egbert B. Gebstadter in *Literary Australian,* Vol. 1, No. 248, pages 22–14; January, 1983.

THE LARGE-SCALE CULTIVATION OF MAMMALIAN CELLS

TISSUE CULTURE: METHODS AND APPLICATIONS. Edited by Paul F. Kruse, Jr., and M. K. Patterson, Jr. Academic Press, 1973.

CELL CULTURE. Edited by William B. Jacoby and Ira H. Pastan. Academic Press, 1979.

OPTIMIZING CULTURE CONDITIONS FOR THE PRODUCTION OF ANIMAL CELLS IN MICROCARRIER CULTURE. J. M. Clark and M. D. Hirtenstein in *Annals of the New York Academy of Sciences,* Vol. 369, pages 33–46; 1981.

MACHINES THAT WALK

ANIMALS IN MOTION. Eadweard Muybridge. Dover Publications, Inc., 1957.

USING BODY SIZE TO UNDERSTAND THE STRUCTURAL DESIGN OF ANIMALS: QUADRUPED LOCOMOTION. T. A. McMahon in *Journal of Applied Physiology,* Vol. 39, No. 4, pages 619–627; 1975.

ADAPTIVE LOCOMOTION OF A MULTI-LEGGED ROBOT OVER ROUGH TERRAIN. R. B. McGhee and G. I. Iswandhi in *IEEE Transactions on Systems, Man, and Cybernetics,* Vol. SMC-9, No. 4, pages 176–182; 1979.

DYNAMIC STABILITY AND RESONANCE IN A ONE-LEGGED HOPPING MACHINE. M. H. Raibert in *Proceedings of the 4th Symposium on Theory and Practice of Robots and Manipulators,* edited by A. Morecki, G. Bianchi and K. Kedzior. Elsevier Scientific Publishing Co., 1981.

A WALKING ROBOT. Ivan E. Sutherland. The Marcian Chronicles, Inc., Pittsburgh, 1982.

THE HAIR CELLS OF THE INNER EAR

THE AUDITORY PERIPHERY: BIOPHYSICS AND PHYSIOLOGY. Peter Dallos. Academic Press, 1973.

SENSITIVITY, POLARITY, AND CONDUCTANCE CHANGE IN THE RESPONSE OF VERTEBRATE HAIR CELLS TO CONTROLLED MECHANICAL STIMULI. A. J. Hudspeth and D. P. Corey in *Proceedings of the National Academy of Sciences of the United States of America,* Vol. 74, No. 6, pages 2407–2411; June, 1977.

THE FOOTPRINTS OF EXTINCT ANIMALS

ICHNIA TETRAPODORUM. O. Kuhn in *Fossilium Catalogus I: Animalia,* edited by F. Westphal. W. Junk, The Hague, 1963.

DIE FOSSILIEN CAURIERFÄHRTEN. Hartmut Haubold. Die Neue Brehm-Bücherei, Ziemsen, Wittenberg Lutherstadt, 1974.

FOSSIL TRACKS AND IMPRESSIONS OF VERTEBRATES. William A. S. Sarjeant in *The Study of Trace Fossils: A Synthesis of Principles, Problems, and Procedures in Ichnology,* edited by Robert W. Frey. Springer-Verlag, 1975.

NMR SPECTROSCOPY OF LIVING CELLS

CELLULAR APPLICATIONS OF ^{31}P AND ^{13}C NUCLEAR MAGNETIC RESONANCE. R. G. Shulman, T. R. Brown, K. Ugurbil, S. Ogawa, S. M. Cohen and J. A. den Hollander in *Science,* Vol. 205, No. 4402, pages 160–166; July 13, 1979.

MAPPING OF METABOLITES IN WHOLE ANIMALS BY ^{31}P NMR USING SURFACE COILS. Joseph J. H. Ackerman, Thomas H. Grove, Gordon G. Wong, David G. Gadian and George K. Radda in *Nature,* Vol. 283, No. 5743, pages 167–170; January 10, 1980.

EXAMINATION OF A CASE OF SUSPECTED MCARDLE'S SYNDROME BY ^{31}P NUCLEAR MAGNETIC RESONANCE. Brian D. Ross, George K. Radda, David G. Gadian, Graeme Rocker, Margaret Esiri and James Falconer-Smith in *The New England Journal of Medicine,* Vol. 304, No. 22, pages 1338–1342; May 28, 1981.

IN VIVO CARBON-13 NUCLEAR MAGNETIC RESONANCE STUDIES OF MAMMALS. J. R. Alger, L. O. Sillerud, K. L. Behar, R. J. Gillies, R. G. Shulman, R. E. Gordon, D. Shaw and P. E. Hanley in *Science,* Vol. 214, No. 4521, pages 660–662; November 6, 1981.

NUCLEAR MAGNETIC RESONANCE AND ITS APPLICATIONS TO LIVING SYSTEMS. David G. Gadian. Oxford University Press, 1982.

THE PHYSICS OF ORGAN PIPES

SOUNDING MECHANISM OF THE FLUTE AND ORGAN PIPE. John W. Coltman in *The Journal of the Acoustical Society of America,* Vol. 44, No. 4, pages 983–992; October, 1968.

ON THE MECHANISM OF SOUND PRODUCTION IN ORGAN PIPES. Samuel A. Elder in *The Journal of the Acoustical Society of America,* Vol. 54, No. 6, pages 1554–1564; December, 1973.

SOUND PRODUCTION BY ORGAN FLUE PIPES. N. H. Fletcher in *The Journal of the Acoustical Society of America,* Vol. 60, No. 4, pages 926–936; October, 1976.

WAVE PROPAGATION ON TURBULENT JETS. S. Thwaites and N. H. Fletcher in *Acustica,* Vol. 45, No. 3, pages 175–179; July, 1980.

HARMONIC GENERATION IN ORGAN PIPES, RECORDERS AND FLUTES. N. H. Fletcher and Lorna M. Douglas in *The Journal of the Acoustical Society of America,* Vol. 68, No. 3, pages 767–771; September, 1980.

WAVE PROPAGATION ON TURBULENT JETS, II: GROWTH. S. Thwaites and N. H. Fletcher in *Acustica,* Vol. 51, No. 1, pages 44–49; June, 1982.

THE ORIGIN OF THE COSMIC X-RAY BACKGROUND

THE LUMINOSITY OF SERENDIPITOUS X-RAY QSOS. Bruce Margon, Gary A. Chanan and Ronald A. Downer in *The Astrophysical Journal,* Vol. 253, No. 1, Part 2, pages L7-L11; February 1, 1982.

THE MASS PRODUCTION OF IRON CASTINGS IN ANCIENT CHINA

THE DEVELOPMENT OF IRON AND STEEL TECHNOLOGY IN CHINA. Joseph Needham. The Newcomen Society, 1958.

T'IEN KUNG K'AI-WU (CHINESE TECHNOLOGY IN THE 17TH CENTURY). Sung Ying-Hsing, translated by E-Tu Zen Sun and Shiou-Chuan Sun. The Pennsylvania State University, 1966.

A HISTORY OF METALLURGY. R. F. Tylecote. The Metals Society, 1976.

THE AMATEUR SCIENTIST

LISTENING TO THE DOPPLER SHIFT OF VISIBLE LIGHT. R. E. Crandall and E. H. Wishnow in *American Journal of Physics,* Vol. 49, No. 5, pages 477–478; May, 1981.

Marygrove College Library
Detroit, Michigan 48221
PLEASE DO NOT REMOVE

SCIENTIFIC AMERICAN

SYNTHETIC VACCINES

$2.50

February 1983

0 74820 08715 02

Introducing the
Reshaped. Refined
And every

Get it together — Buckle up

Established 1845 **SCIENTIFIC AMERICAN** February 1983 Volume 248 Number 2

ARTICLES

45 **THE FUTURE OF AMERICAN AGRICULTURE, by Sandra S. Batie and Robert G. Healy** The main factors are the supply of land, water and energy and the demand for exports.

54 **THE LATTICE THEORY OF QUARK CONFINEMENT, by Claudio Rebbi**
The force between quarks is simulated by imposing a lattice on the structure of space and time.

66 **SYNTHETIC VACCINES, by Richard A. Lerner**
Chains of amino acids made in the laboratory mimic viral proteins and give rise to antibodies.

85 **THE OPTICAL COMPUTER, by Eitan Abraham, Colin T. Seaton and S. Desmond Smith**
An optical analogue of the transistor is the crucial element of a computer based on beams of light.

94 **HIDDEN VISUAL PROCESSES, by Jeremy M. Wolfe**
The visual system includes subsystems whose operation is normally not apparent to the perceiver.

104 **THE ACTIVE SOLAR CORONA, by Richard Wolfson**
The dynamics of the sun's outer atmosphere reflect the interaction of matter and magnetic fields.

120 **DART-POISON FROGS, by Charles W. Myers and John W. Daly**
In Colombia, Indian hunters poison blowgun darts with the highly toxic alkaloids of small frogs.

134 **FAN VAULTING, by Walter C. Leedy, Jr.**
In the 14th century English masons converted the Gothic arch into a two-dimensional curved shell.

DEPARTMENTS

6 LETTERS
11 50 AND 100 YEARS AGO
12 THE AUTHORS
14 METAMAGICAL THEMAS
30 BOOKS
76 SCIENCE AND THE CITIZEN
146 THE AMATEUR SCIENTIST
154 BIBLIOGRAPHY

BOARD OF EDITORS Gerard Piel (Publisher), Dennis Flanagan (Editor), Brian P. Hayes (Associate Editor), Philip Morrison (Book Editor), Francis Bello, John M. Benditt, Peter G. Brown, Michael Feirtag, Diana Lutz, Jonathan B. Piel, John Purcell, James T. Rogers, Armand Schwab, Jr., Joseph Wisnovsky

ART DEPARTMENT Samuel L. Howard (Art Director), Steven R. Black (Assistant Art Director), Ilil Arbel, Edward Bell

PRODUCTION DEPARTMENT Richard Sasso (Production Manager), Carol Hansen and Leo J. Petruzzi (Assistants to the Production Manager), Carol Eisler (Senior Production Associate), Karen O'Connor (Assistant Production Manager), Carol Albert, Lori Mogol, Martin O. K. Paul, Julio E. Xavier

COPY DEPARTMENT Sally Porter Jenks (Copy Chief), Debra Q. Bennett, Mary Knight, Dorothy R. Patterson

GENERAL MANAGER George S. Conn

ADVERTISING DIRECTOR C. John Kirby

CIRCULATION MANAGER William H. Yokel

SECRETARY Arlene Wright

SCIENTIFIC AMERICAN (ISSN 0036-8733), PUBLISHED MONTHLY BY SCIENTIFIC AMERICAN, INC., 415 MADISON AVENUE, NEW YORK, N.Y. 10017. COPYRIGHT © 1983 BY SCIENTIFIC AMERICAN, INC. ALL RIGHTS RESERVED. PRINTED IN THE U.S.A. NO PART OF THIS ISSUE MAY BE REPRODUCED BY ANY MECHANICAL, PHOTOGRAPHIC OR ELECTRONIC PROCESS, OR IN THE FORM OF A PHONOGRAPHIC RECORDING, NOR MAY IT BE STORED IN A RETRIEVAL SYSTEM, TRANSMITTED OR OTHERWISE COPIED FOR PUBLIC OR PRIVATE USE WITHOUT WRITTEN PERMISSION OF THE PUBLISHER. SECOND-CLASS POSTAGE PAID AT NEW YORK, N.Y., AND AT ADDITIONAL MAILING OFFICES. AUTHORIZED AS SECOND-CLASS MAIL BY THE POST OFFICE DEPARTMENT, OTTAWA, CANADA, AND FOR PAYMENT OF POSTAGE IN CASH. SUBSCRIPTION RATE: $21 PER YEAR IN THE U.S. AND ITS POSSESSIONS; $27 PER YEAR IN ALL OTHER COUNTRIES. POSTMASTER: SEND ADDRESS CHANGES TO SCIENTIFIC AMERICAN, 415 MADISON AVENUE, NEW YORK, N.Y. 10017.

SCIENTIFIC AMERICAN

CORRESPONDENCE

Offprints of more than 1,000 selected articles from earlier issues of this magazine, listed in an annual catalogue, are available at 75 cents each. Correspondence, orders and requests for the catalogue should be addressed to W. H. Freeman and Company, 660 Market Street, San Francisco, Calif. 94104. Offprints adopted for classroom use may be ordered direct or through a college bookstore. Student sets of 10 or more Offprints are collated by the publisher and are delivered as sets to bookstores.

Photocopying rights are hereby granted by Scientific American, Inc., to libraries and others registered with the Copyright Clearance Center (CCC) to photocopy articles in this issue of SCIENTIFIC AMERICAN for the flat fee of $2 per copy of each article or any part thereof. Such clearance does not extend to the photocopying of articles for promotion or other commercial purposes. Correspondence and payment should be addressed to Copyright Clearance Center, Inc., 21 Congress Street, Salem, Mass. 01970. Specify CCC Reference Number ISSN 0036–8733/83. $2.00 + 0.00.

Editorial correspondence should be addressed to The Editors, SCIENTIFIC AMERICAN, 415 Madison Avenue, New York, N.Y. 10017. Manuscripts are submitted at the authors' risk and will not be returned unless they are accompanied by postage.

Advertising correspondence should be addressed to C. John Kirby, Advertising Director, SCIENTIFIC AMERICAN, 415 Madison Avenue, New York, N.Y. 10017.

Subscription correspondence should be addressed to Subscription Manager, SCIENTIFIC AMERICAN, P.O. Box 5969, New York, N.Y. 10017. The date of the last issue on your subscription is shown in the upper right-hand corner of each month's mailing label. For change of address notify us at least four weeks in advance. Please send your old address (if convenient, on a mailing label of a recent issue) as well as the new one.

Name

New Address

Street

City

State and ZIP

Old Address

Street

City

State and ZIP

THE COVER

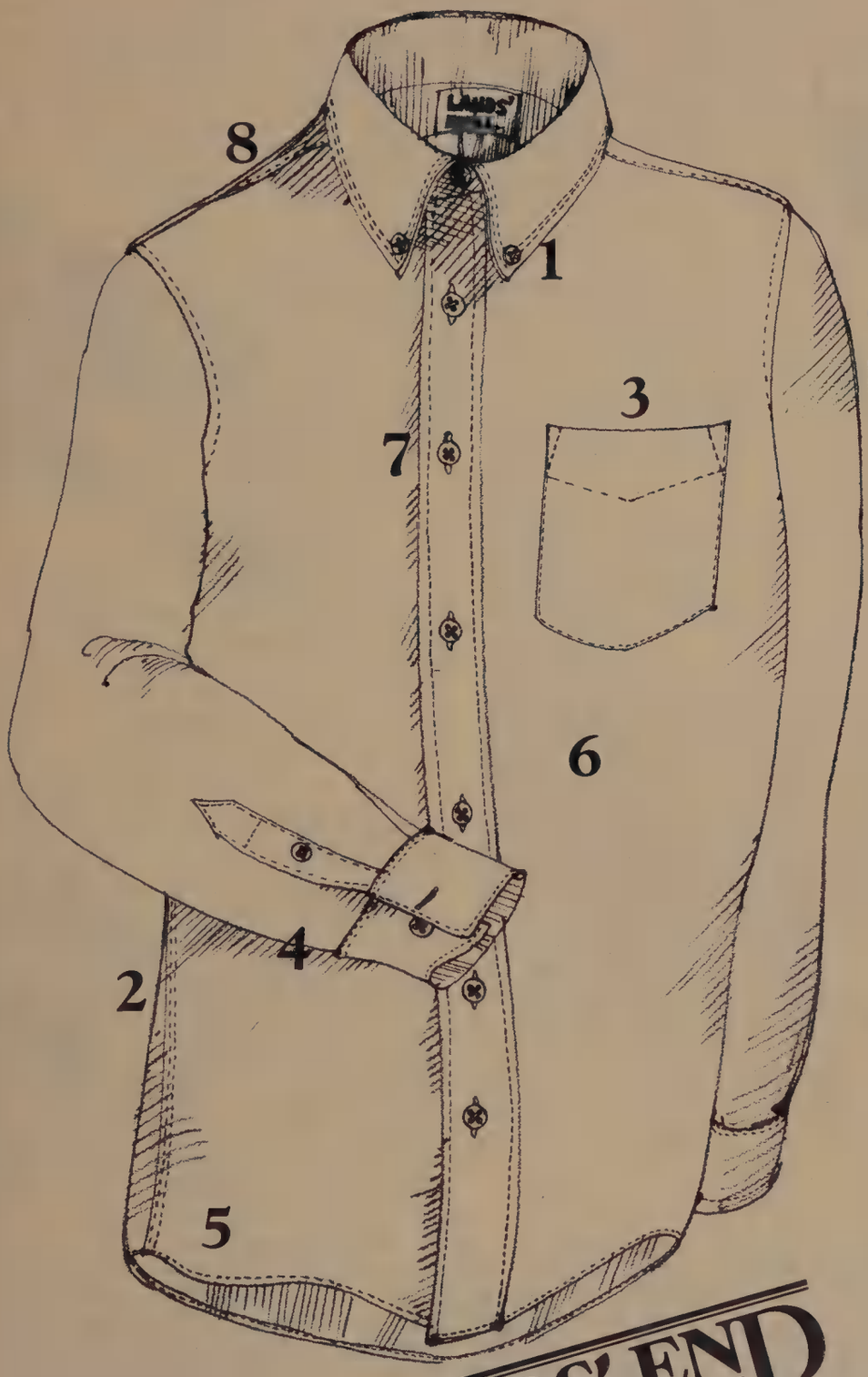
The illustration on the cover shows (in blue) the polyhedral structure of the capsid, or protein shell, of the tomato bushy-stunt virus. The capsid is composed of 180 copies of a protein, with three copies occupying each of the 60 triangular faces of the shell; 12 copies of the surface domain of the protein are depicted in orange. The illustration is based on a computer display generated by Arthur J. Olson to demonstrate a step in the design of a synthetic vaccine (see "Synthetic Vaccines," by Richard A. Lerner, page 66). Whereas a vaccine directed against viral infection ordinarily contains intact viruses, a synthetic vaccine contains only a peptide (a short segment of a protein chain) synthesized in the laboratory to mimic part of a viral surface protein. One way to ensure that the peptide mimics a site exposed on the surface, and therefore is likely to elicit antibodies that recognize the intact virus, is to exploit computer-graphics methods developed by T. J. O'Donnell of Abbott Laboratories and Michael L. Connolly and Olson of the Research Institute of Scripps Clinic.

THE ILLUSTRATIONS

Cover illustration by Gabor Kiss

Page	Source
46–53	Andrew Tomko
55–65	Ilil Arbel
67	Arthur J. Olson, Research Institute of Scripps Clinic
68–69	Patricia J. Wynne
70–73	Arthur J. Olson, Research Institute of Scripps Clinic
74	Patricia J. Wynne
86–93	Walken Graphics
95	R. F. Bonifield, courtesy of the Granger Collection
96–102	George V. Kelvin
105	High Altitude Observatory
106	Jay M. Pasachoff, Williams College (*top*); High Altitude Observatory (*bottom*)
108	High Altitude Observatory
110	American Science & Engineering, Inc.
111	Goddard Space Flight Center (*top*), High Altitude Observatory (*bottom*)
112	Naval Research Laboratory
114–119	Allen Beechel
121	David M. Dennis
124	Patricia J. Wynne
125	Charles W. Myers, American Museum of Natural History
127	Lois W. Tice, National Institute of Arthritis, Diabetes, and Digestive and Kidney Diseases
128	Patricia J. Wynne
132	Enid Kotschnig
134	John Moss, Black Star
136–139	Alan D. Iselin
142	Walter C. Leedy, Jr., Cleveland State University (*top*); Alan D. Iselin (*bottom*)
143	Alan D. Iselin
144	John Moss, Black Star
146–152	Michael Goodman

We set out to make the world's best buttondown. This one comes close. At $24.



This is the Hyde Park—the latest addition to our impressive Oxford Collection, featuring both shirts of 100% cotton and our Lands' End reverse cotton blends.

Check it out feature for feature, beginning with the knowledge that it's made of imported 100% cotton Oxford. Heavier. More densely woven for a nicer drape. It launders better, resists wrinkles; best of all, it wears longer than normal.

For those of you interested in more specifics, we've provided this step-by-step "tour" of the shirt—available in pink, ecru, blue, maize, helio and white in our catalog.

1. **Double-track stitching highlights the softly-rolled collar, keeps it neater with a tie.**
2. **The shirt is single-needle stitched, with strong double-needle side seams.**
3. **Note the placement and detailing of the left breast pocket.**
4. **Barrel cuffs have a long, buttoned placket.**
5. **The shirt has extra long tails, so it stays neatly tucked in.**
6. **The cut is full, for comfort. Yet civilized, too. Tapered just enough to avoid being sloppy.**
7. **Note that it has a generous 7-button front.**
8. **In back, a box pleat with hanging loop adds function and style.**

Why make so much of a single shirt? We may have told you more than you ever wanted to know about a shirt. But only to make a point about the Lands' End philosophy of doing business. It is a simple philosophy really:

First, *quality*. Then, *price*. And always, always *service*.

A quality item at a reasonable price represents a Lands' End value. Anything less is someone else's ballgame—not ours. What's more, every item we offer you—from soft luggage to sweaters to snow wear to shoes—is unconditionally guaranteed.

We don't ask you to trust us, just try us. Mail the coupon for a free catalog. Better still call our toll-free 800 number, 800-356-4444. 24 hours a day. Except in Wisconsin, 608-935-2788.

LANDS' END
DIRECT MERCHANTS

of fine wool and cotton sweaters, Oxford buttondown shirts, traditional dress clothing, snow wear, deck wear, original Lands' End soft luggage and a multitude of other quality goods from around the world.

☐ Please send free catalog.
Lands' End Dept. Q-08 Dodgeville, WI 53533

Name _____

Address _____

City _____

State _____ Zip _____

Or call Toll-free
800-356-4444
(Except Wisconsin call 608-935-2788)

NordicTrack

Jarless Total Body Cardiovascular Exerciser Duplicates X-C Skiing for the Best Motion in Fitness

The enjoyable sport of cross-country skiing is often cited by physiologists as the most perfect form of cardiovascular exercise for both men and women. Its smooth, fluid, total body motion uniformly exercises more muscles and higher heart rates seem easier to attain than when jogging or cycling. NordicTrack closely simulates the pleasant X-C skiing motion and provides the same cardiovascular endurance-building benefits—right in the convenience of your home. Makes a year round, consistent exercise program easily attainable. Eliminates the usual barriers of time, weather, chance of injury, etc. Also highly effective for weight control.

Better Than Running
NordicTrack gives you a more complete work out—conditions both upper body and lower body muscles at the same time. Fluid, jarless motion will not cause joint or back pains as jogging or running often does.

More Effective Than Exercise Bikes
NordicTrack's stand-up skiing motion more uniformly exercises large leg muscles and also adds important upper body exercise. Higher pulse rates, necessary for building fitness, seem easier to attain because the work is shared by more muscle mass. The NordicTrack also keeps more muscles in tone.

Even Better Than Swimming
NordicTrack more effectively exercises the largest muscles in the body, those located in the legs and buttocks. These muscles' main function is to lift and propel the body in the standing position. However, when swimming, the body is supported by the water, thus preventing these muscles from being effectively exercised. The stand up exercising position on the NordicTrack much more effectively exercises these muscles.

A Proven, High Quality Durable Product
Ruggedly constructed, NordicTrack is quiet, motorless and folds compactly for storage. Separately adjustable arm and leg resistances. NordicTrack is in its 6th year of production and is used in thousands of homes and many institutions. We manufacture and sell direct. One year warrantee, 15 day trial period with return privilege.

Call or Write for FREE BROCHURE
TOLL FREE 1-800-328-5888
8 AM-5 PM MON.-FRI.
PSI 124 F Columbia ct. Chaska, MN 55318
Minnesota 612-448-6987

LETTERS

Sirs:

I must congratulate Randall Forsberg on her objective appraisal of the state of the nuclear arms race and its concomitant dangers ["A Bilateral Nuclear-Weapon Freeze," by Randall Forsberg; SCIENTIFIC AMERICAN, November, 1982]. For me the article has provided the most convincing case for a nuclear freeze yet seen in print.... To Europeans, who fear the outbreak of nuclear war on their soil as keenly as, if not more than, citizens of the U.S., and particularly in view of the weapon escalation planned by NATO and the British government for next year, this article has been most timely.

Whilst a bilateral nuclear freeze would be strongly welcomed, Ms. Forsberg did not mention a complementary move that could be tied to a freeze and that could considerably improve the security of all nations: the establishment of a European battlefield-nuclear-weapon-free zone (BNWFZ) on both sides of the Iron Curtain, as has been recommended by the Independent Commission on Disarmament and Security Issues under the chairmanship of Olof Palme (*Common Security: A Blueprint for Survival,* Simon and Schuster, 1982). The existing dangers of current battlefield nuclear weapons have been clearly stated in the Palme report: "Because of their deployment in forward areas battlefield nuclear weapons run the risk of being overrun early in an armed conflict. Maintaining command and control over such weapons in 'the fog of war' would be difficult. Pressures for delegation of authority to use nuclear weapons to local commanders and for their early use would be strong. The danger of crossing the nuclear threshold and of further escalation could become acute."

GRAEME G. WILKINSON

University of Reading
Reading, England

Sirs:

I was interested in the "default assumption" story about the son whose mother was a surgeon ["Metamagical Themas," by Douglas R. Hofstadter; SCIENTIFIC AMERICAN, November, 1982]. When I translated the story into Latvian, my native language, the default assumption failed. The rules of grammar revealed the surgeon to be female. I did construct some default-assumption stories in the language and tried them on my countrypersons. My reward was reprimands for improper grammar, distorted syntax, wordplay and tricksterism.

There are compensations. Consider the plight of two Latvian couples, Freda and Fredis Ozols and Vila and Vilis Liepa. Fredis was pleased with his surname Ozols, the oak tree, a symbol of manly strength and fortitude. Freda, like many other wives and daughters, was not. There was a national debate on the question, and eventually women were allowed to feminize their surnames. Thus Freda Ozols became Freda Ozola. Fredis did not object; Ozola is also the genitive case of Ozols, making Freda Fredis' property.

Meanwhile Vilis Liepa led a compromising life with a feminine-gender surname, the linden tree, a symbol of womanly propriety and virtue. Vila was pleased but secretly envied Freda, who could differentiate her surname from her husband's. Vila took the genitive strategy and became Vila Liepas, which also masculinizes the surname. It is odd, grumbled Vilis; if I were a woman, I could have a proper man's name.

KARLIS REKEVICS

Seattle, Wash.

Sirs:

The mention of Mersenne numbers in "The Search for Prime Numbers," by Carl Pomerance [SCIENTIFIC AMERICAN, December, 1982], brings to mind an anecdote told by E. T. Bell in *Mathematics: Queen and Servant of the Sciences.* Marin Mersenne had claimed that $2^{67} - 1$ was a prime number, but in 1903 F. N. Cole proved it was composite. To quote Bell:

"When I asked Cole in 1911 how long it had taken him to crack M_{67}, he said 'Three years of Sundays.' But this, though interesting, is not the history. At the October 1903 meeting in New York of the American Mathematical Society, Cole had a paper on the program with the modest title 'On the factorization of large numbers.' When the chairman called on him for his paper, Cole—who was always a man of very few words—walked to the board and, saying nothing, proceeded to chalk up the arithmetic for raising 2 to the sixty-seventh power. Then he carefully subtracted 1. Without a word he moved over to a clear space on the board and multiplied out, by longhand, $193,707,721 \times 761,838,257,287$. The two calculations agreed. Mersenne's conjecture—if such it was—vanished into the limbo of mathematical mythology. For the first and only time on record, an audience of the American Mathematical Society vigorously applauded the author of a paper delivered before it. Cole took his seat without having uttered a word. Nobody asked him a question."

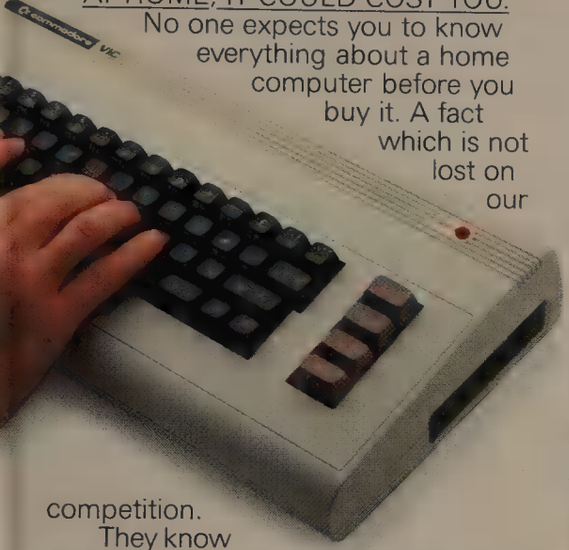
ROBERT C. GEBHARDT

Hopatcong, N.J.

HOME IS THE LAST PLACE YOU SHOULD LEARN ABOUT A HOME COMPUTER.

WANT TO LEARN SOMETHING ABOUT HOME COMPUTERS? HERE, IT'S FREE. AT HOME, IT COULD COST YOU.

No one expects you to know everything about a home computer before you buy it. A fact which is not lost on our competition.

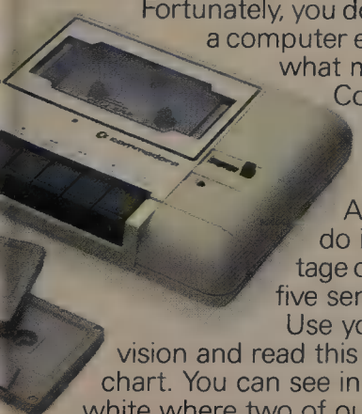


They know that an impressively low price can divert your attention from some depressingly cheap features. So that you won't know what you may be missing with their home computer until after it's been in your home for a while.

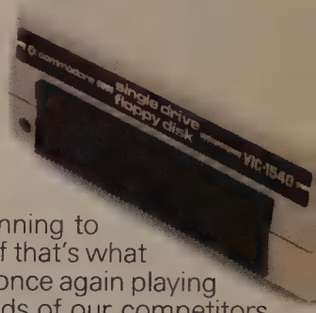
At which point, naturally, it'll cost you to change your mind.

IT'S EASY TO TELL THE DIFFERENCE.

Fortunately, you don't have to be a computer engineer to tell what makes the Commodore VIC 20™ superior to the competition. All you have to do is take advantage of three of your five senses.



Use your sense of vision and read this comparison chart. You can see in black and white where two of our major competitors have skimped. Use your sense of touch in the store. You'll feel the VIC 20's superiority immediately. It feels a lot more expensive than it is.



If these two senses don't convince you that the VIC 20 offers more for the money than any other home computer, simply rely on common sense.

NOW THAT YOU KNOW HOW EASY A COMMODORE HOME COMPUTER IS TO OWN, FIND OUT HOW EASY IT IS TO EXPAND.

One thing about home computers that you're bound to discover at home is that, once you learn what they can do, you'll want them to do more and more. To do this, you may need accessories called peripherals. These let you get more out of a home computer by letting you put more into it.

They include items like cassette recorders and disk drives to input data, modems for telecomputing and printers. And all VIC 20 peripherals are fully compatible with the powerful Commodore 64™ personal computer.

PLAN AHEAD.

When you start looking at your first home computer, you may think it's too early to start planning to add peripherals. If that's what you think, you're once again playing right into the hands of our competitors.

Because once they've gotten you to buy their home computer, for what seems to be a reasonable price, they have you hooked on their system.

The costs of which, if you'll examine the chart below, can really start getting unreasonable. For example, while these computers may seem to be close to the same price to start, an expanded system can cost you twice as much with TI or Atari as with the Commodore VIC 20 or Commodore 64.

EXPANSION COSTS	VIC 20™ or COMMODORE 64™	TI99/4A®	ATARI 400®
BASIC	Included	Included	$59.95
Peripheral Expansion System	Not Necessary	$249.95	Not Necessary
Disk Drive	$399.00	399.95	599.95
Disk Controller Card	Included	249.95	Included
Modem	109.95	224.95	199.95
Modem Interface	Included	174.95	219.95
TOTAL	$508.95	$1299.75	$1079.80

Manufacturer's suggested list prices: Prices per TI June-December 1982 U.S. Consumer Products Suggested Price List. Atari prices effective July 1, 1982 Suggested Retail Price List.

THINK OF IT AS BUYING A TOASTER.

It's easy to fill up a computer ad with RAM's and ROM's, numbers and technical jargon. But when it comes right down to it, buying a home computer is just like buying anything else. It's important to know just what you're getting for your hard-earned money.

And we hope we've accomplished that here by telling you about the cost of expanding your Commodore VIC 20 or Commodore 64 computer.

COMPUTER FEATURES	VIC 20	TI 99/4A	ATARI 400
Typewriter Keys	Yes	Yes	No
Typewriter Feel	Yes	No	No
Color Control Keys	Yes	No	No
Graphics on Keys	Yes	No	No
Reverse Letters	Yes	No	Yes
Programmable Function Keys	Yes	No	No
Works with TV or Monitor	Yes	Yes	No
True Lower Case Letters	Yes	No	Yes
DISK FEATURES			
Capacity	170K	90K	88K

commodore COMPUTER

U.S.A.—P.O. Box 500, Conshohocken, PA 19428; Canada—3370 Pharmacy Ave., Agincourt, Ontario, Canada M1W2K4.

Quattro Turbo

4000 Sports Sedan or Diesel

Gas model shown here

5000 Turbo Diesel

GT Coupe

Audi offers more high-performance German luxury cars than anyone. Surprised?

Audi In a sea of newly minted "performance" cars, automakers are suddenly extolling the pleasures of "varroom."

Performance is, once more, "in."

At Audi, it was never "out."

After decades of refining performance automobiles, our engineers in Germany find all this amusing.

To those discerning car buyers, the most sophisticated high-performance luxury cars continue to come from Germany.

The most complete line of which comes from Audi, for 1983. The recent addition of the 5000 Turbo Diesel, the Quattro Turbo and GT Coupe illustrates Audi's commitment to performance.

Without hyperbole, we can say that no luxury Sports Coupe can compete with our Quattro Turbo.

Ask those rally drivers who have tried to, in America and in Europe. The Quattro Turbo's powerhouse tandem of all-wheel drive and turbocharged five-cylinder engine has led to eight straight victories in SCCA competition in the United States—following one stunning victory after another in Europe.

Turbocharged engine power is also readily available in our 5000 flagship line. The 5000 Turbo Gas and the 5000 Turbo Diesel. Two full-sized, high-performance German luxury cars with turbo exhilaration. A rare blending of elegance, power and remarkably quick reflexive handling.

In a mere 7.5 seconds, the 3000-pound 5000 Turbo can bring you comfortably to 50 miles an hour. Even the Turbo Diesel is a top performer with its 2.0-liter, five-cylinder power plant.

When we introduced our GT Coupe, it was named, without question, by the editors of *Road & Track* "Best Sports Coupe For The 80's."

We weren't surprised. It is one of the few genuine Grand Touring Coupes. And the only front-wheel drive, five-cylinder Coupe in the world. Aerodynamically designed, it thrives on tight, hairpin curves.

Some of the most enjoyable performance driving you'll ever do will be in our 4000 Sports Sedan. It is extremely agile, the steering very light. Front-wheel drive. Four-cylinder, 1.7-liter engine.

Available as a two-door, four-door, or '82 diesel model, the 4000s offer many exciting features as standard equipment and have a low starting base price of only $10,205.*

No matter which Audi you buy, you get the high level of luxury you expect from motor cars in this class. Because, here in Ingolstadt, Germany, we engineer thoughtful luxuries into every Audi. For example, a few of the luxurious appointments available in certain Audi models include: air conditioning, power windows, central door/trunk locking system, sumptuous seats (genuine leather available) and AM/FM stereo/cassette radio.

Simply stated, in a luxury Audi, performance means something today. Because it always has.

For your nearest Porsche Audi dealer or details on the Audi Delivery in Europe Program, call toll-free (800) 447-4700. In Illinois, (800) 322-4400.

© 1983 Porsche Audi

PORSCHE + AUDI

Audi: the art of engineering.

Use "est. mpg" for comparison. Mpg and range vary with speed, trip length, weather. Actual hwy. mpg and hwy. range will be less. *Use mfr's sugg. retail price. Title, taxes, transp., registration and dealer prep add'l.

A carat or more. A rare diamond. For that rare individual who never doubted you'd make it.

The one and a half carat diamond necklace shown below is enlarged for detail.

A diamond of a carat or more. There's only one in a million.

Every diamond is rare. But of all diamonds found, a solitaire of a carat or more is only one in a million.

And, like love, becomes more precious with time.

A miracle among miracles. Born from the earth. Reborn on a woman.

The extraordinary diamond of a carat or more.

Show the world you wouldn't have made it without her.

A diamond is forever. De Beers.

50 AND 100 YEARS AGO

SCIENTIFIC AMERICAN

FEBRUARY, 1933: "A million dollars a night! Dr. Irving Langmuir, who has just received the Nobel award for chemistry, made possible this saving to the American people in electricity costs by his development of the high-intensity incandescent lamp, which contains small quantities of either nitrogen or argon. He is only the second American chemist to have been honored by the Swedish Academy in 31 years. Growing out of his work with the incandescent lamp came momentous discoveries affecting the then new and immature vacuum tube. Dr. Langmuir's work with vacuum tubes contributed largely to the vast development of modern radio broadcasting. Equally important was his discovery of the atomic-hydrogen method of arc welding. He has been particularly interested in the mechanism of chemical reactions taking place on the surfaces of solids. He is now engaged in working out the laws according to which atoms and molecules distribute themselves over surfaces to form monatomic layers—laws of great importance in understanding many simple phenomena."

"Stars that appear reddish in color have been thought to be very far away, but now Dr. Joel Stebbins of the Mount Wilson Observatory reports to the National Academy of Sciences meeting at the University of Michigan that some red stars located in the Milky Way are four times closer to us than has been estimated. A layer of 'cosmic dust' that splits the Milky Way makes the stars on the far side appear red, just as our sun appears red at sunset, when its rays come to the observer through a longer distance of the dust-filled atmosphere of our planet."

"The rocket may someday become a very dangerous military weapon. In fact, many authorities think that its utilization in warfare is likely to come much sooner than its application in aerial navigation, and they visualize a powerful rocket guided by wireless and capable of destruction at long range. Perhaps one reason the Germans are so interested in rocket development is that they are debarred from the construction of military aircraft. At any rate, more rocket experimentation is carried on in Germany than in all other countries put together. The latest German rocket design is that of a young engineer named Tilling, of Osnabrück. Tilling recently gave a striking demonstration at the Tempelhof aerodrome in Berlin. The Tilling rocket is made entirely of aluminum and consists of a central shell to which are attached four long tail fins. The over-all length is about nine feet. The rocket motor is inserted into the shell between the fin roots; this motor is filled with a special powder that is relatively slow-burning. As Tilling wished his craft to alight within the aerodrome, he only employed a 12-pound charge in his experiment. The powder was ignited from a distance by means of an electric spark, and the reaction of the exhaust gases drove the rocket upward to a height of about 2,600 feet. Two wings were hidden in the hollow fins until the rocket had reached its maximum altitude. Then a simple automatically acting hydraulic device gradually moved the wings out from their sheaths. The rocket, now converted into a glider, made its descent in a fairly tight spiral. The landing was only 400 yards from the starting point."

FEBRUARY, 1883: "At a recent meeting of the Académie de Médecine, M. Bouley communicated, in the name of M. Pasteur, a series of conclusions regarding rabies at which the distinguished investigator has arrived. The first was to enunciate the familiar truths that the dumb madness and the furious madness, and in short all varieties of rabies, are caused by the same virus. Mention was also made of the cases of three dogs inoculated in 1881: two quickly died from rabies; the third, after manifesting the early systems, recovered. The latter animal was inoculated by trephining in 1882 on two separate occasions, but without effect. M. Pasteur asserts that he now has four dogs that will not contract rabies, whatever the method of inoculation adopted or the proved virulence of the material employed. These facts he believes to be the first step toward the discovery of a method of the prevention of rabies by its inoculation. He confesses, however, that the end seems to be at present far distant."

"It is simply for lack of muscular power that man cannot fly. There is no combination of wings or arrangements of any kind that will compensate for this fact. Whether man can produce a machine to supplement his own want of force remains to be seen. Such a motor cannot, we think, be driven by steam. It is not impossible that a machine might be made that would be caused to fly by means of a small electric motor run at a very high speed and worked by the aid of a couple of wires from the ground. This, however, would hardly be flying in the true sense of the word. That wings and such things can be made we have no doubt; and experiments enough have been done to prove that, if power enough be available, flight can be achieved. When a machine can be made each pound of which will develop as much energy as each pound of a bird, flying may be possible—not till then."

"The Hudson River ice crop of the season just closed is rated as one of the largest, if not the largest, ever harvested. It amounted to nearly 3,000,000 tons. It is said that not so many new ice houses were built last fall as the previous one; still, those constructed were quite large, and they increase the storing capacity 200,000 tons. The ice gathered ranges in thickness from eight to 20 inches, and, owing to the low condition of the river when it froze over, is as clear as crystal. Notwithstanding the cry of short crop last winter, about 100,000 tons remained in ice houses unsold at the beginning of the year. Nearly 50,000 tons of this was ice gathered in 1880, which had been held for a good market."

"It is gratifying to know that the Western Union Telegraph Company have decided to begin in earnest the work of putting all the wires of the company in New York underground. The first underground main will extend from the company's building at Broadway and Dey Street, up Broadway to 23rd Street. The wires, it is said, will be laid in a tube large enough to contain 200 wires. The right to use the streets for this purpose was secured a few years ago. The promise is that if the underground system adopted works well, it will be extended to all parts of the city."

"Recently newspapers in France announced that the police of Bordeaux had forbidden the exhibition of the automaton Az Rah, one of the attractions of the Exhibition Theater, because it had been discovered that the manikin was set in motion not by mechanical arrangements but by a youth of 18 years enclosed within a cavity behind the wheelwork, and whose health was gravely compromised by this daily torture. This automaton recalls the famous Turkish chess player that was constructed in Hungary by Baron Kempelen in 1769 and exhibited in Germany, Russia, France, England and America without the public's succeeding in ascertaining its mechanism. In 1819 and 1820 a man named Melzer showed it anew in England. Robert Houdin saw it in 1844 at the house of a mechanician of Belleville, named Cronior. Since then its fate has been unknown, and it is very probable the Az Rah of Bordeaux is nothing other than the Turk of Vienna."

THE AUTHORS

SANDRA S. BATIE and ROBERT G. HEALY ("The Future of American Agriculture") are economists with a special interest in natural resources and land-use planning. Batie is associate professor of agricultural economics at Virginia Polytechnic Institute and State University. A graduate of the University of Washington, she has an M.S. and a Ph.D. from Oregon State University. She has been a member of the faculty at Virginia Tech since 1973. Healy is associated with the Conservation Foundation in Washington, D.C. He got his B.A., M.A. and Ph.D. degrees, all in economics, from the University of California at Los Angeles. Before joining the staff of the foundation in 1975 he taught regional planning at Harvard University and U.C.L.A. and was a member of the research staff of the Urban Institute and Resources for the Future. The present article is based in part on an interdisciplinary study sponsored by the Federal Emergency Management Agency and the Conservation Foundation; the study resulted in the book *The Future of American Agriculture as a Strategic Resource,* published by the foundation. Contributors to the book included (besides Batie and Healy) Otto C. Doering III, Kenneth D. Frederick, Jack R. Harlan, Philip M. Raup, Vernon W. Ruttan, Robert H. Shaw and Frederick N. Swader.

CLAUDIO REBBI ("The Lattice Theory of Quark Confinement") is an elementary-particle theorist on the staff of the Brookhaven National Laboratory. Born in Trieste, he studied physics at the University of Torino, where he received the Italian equivalent of a Ph.D. in 1967. He did his postdoctoral work at the California Institute of Technology, taught at the University of Trieste and did research at the European Organization for Nuclear Research (CERN) in Geneva before returning to the U.S. in 1974 as a visiting associate professor at the Massachusetts Institute of Technology. He moved to Brookhaven in 1977. During the academic year 1980–81 he was again at CERN as a visiting scientist. This is Rebbi's second article in SCIENTIFIC AMERICAN; his previous one, "Solitons," appeared in February, 1979.

RICHARD A. LERNER ("Synthetic Vaccines") is chairman of the molecular biology department at the Research Institute of the Scripps Clinic in La Jolla, Calif. He obtained his M.D. from the Stanford University School of Medicine in 1964 and spent the next year at Stanford as a medical intern. He has been affiliated with the Scripps Clinic ever since, except for two years at the Wistar Institute of Anatomy and Biology in Philadelphia in 1968–70. Among his hobbies is writing fiction; his latest published novel is *Epidemic 9,* and he is currently at work on another novel.

EITAN ABRAHAM, COLIN T. SEATON and S. DESMOND SMITH ("The Optical Computer") work together at Heriot-Watt University in Edinburgh. Abraham is a research associate in the physics department, where he lectures on quantum statistics, thermodynamics and computational physics. A native of Israel, he was graduated from the University of Buenos Aires in 1975. He got his Ph.D. from Manchester University in 1979. Seaton received his undergraduate degree in applied physics from Heriot-Watt in 1979 and is currently completing work on his Ph.D. His research interests include nonlinear optical effects in semiconductors and optical switching and signal-processing devices. Smith is professor of physics at Heriot-Watt, where he also serves as dean of the faculty of science. He did his undergraduate work at the University of Bristol and went on to obtain two doctoral degrees: a Ph.D. from the University of Reading (1956) and a D.Sc. from the University of Bristol (1966). Before taking up his present posts at Heriot-Watt in 1970, he worked at the Royal Aircraft Establishment at Farnborough, Imperial College London and the University of Reading. In addition to his academic duties he is chairman of Edinburgh Instruments Ltd., a manufacturer of laser systems. Smith is a fellow of the Royal Society of Edinburgh, the Institute of Physics (which awarded him the C. V. Boys Prize in 1976) and the Royal Society of London. From 1977 to 1980 he was chairman of the quantum electronics division of the European Physical Society.

JEREMY M. WOLFE ("Hidden Visual Processes") teaches introductory psychology at the Massachusetts Institute of Technology and does research in the institute's Center for Cognitive Science. Born in London, he was graduated summa cum laude from Princeton University in 1977. He did his graduate work on binocular vision with Richard Held at M.I.T., getting his Ph.D. in 1981. He writes: "In high school I thought I would become a lawyer or perhaps a rabbi, but I was 'converted' to the experimental psychology of vision by a summer job at Bell Laboratories. I spent my time sitting in a dark room looking at barely visible flashes of colored light and came out determined to do the same to other people." When Wolfe is not working in the classroom or the laboratory, he adds, "I live in Cambridge, Mass., with one wife, two cats (one without a tail) and one six-foot Burmese python."

RICHARD WOLFSON ("The Active Solar Corona") is associate professor of physics at Middlebury College. He spent his undergraduate years at the Massachusetts Institute of Technology and Swarthmore College, receiving a B.A. in philosophy and physics at Swarthmore in 1969. After a brief stint doing research in medical physics at the Massachusetts General Hospital he took a master's degree in environmental studies at the University of Michigan. He subsequently went to Dartmouth College, where he obtained his Ph.D. in physics in 1976. His research interests, he reports, "center on the behavior of gases and plasmas in astrophysical situations and include studies of accretion flows in X-ray sources as well as solar and stellar winds and the solar corona. In 1980–81 I was a visiting scientist at the High Altitude Observatory of the National Center for Atmospheric Research, where I carried out theoretical studies of coronal dynamics. I should like to acknowledge support from the staff of the High Altitude Observatory and from many others who provided photographs."

CHARLES W. MYERS and JOHN W. DALY ("Dart-Poison Frogs") have worked together since 1966, "dividing our time between jungle camps and modern laboratories in a multidisciplinary approach to the biology and biochemistry of poisonous frogs." Myers is chairman of the department of herpetology at the American Museum of Natural History. He received a B.S. at the University of Florida in 1960 and an M.A. at Southern Illinois University in 1962. He interrupted his doctoral studies to spend three years (1964–67) as a visiting scientist at the Gorgas Memorial Laboratory in Panama. He joined the staff of the American Museum in 1968 and got his Ph.D. from the University of Kansas in 1970. Daly is chief of the laboratory of bioorganic chemistry at the National Institute of Arthritis, Diabetes, and Digestive and Kidney Diseases. He obtained a B.S. in 1954 and an M.A. in 1955 at Oregon State College and a Ph.D. in chemistry from Stanford University in 1958.

WALTER C. LEEDY, JR. ("Fan Vaulting"), is an architectural historian and medievalist at Cleveland State University. He majored in architecture as an undergraduate and graduate student at the University of Michigan, taking time out to spend a year as a Kościuszko Foundation Scholar in Poland in 1966–67. He was awarded his Ph.D. in art history in 1972 by the Courtauld Institute of Art of the University of London.

explore the world below

Oceanus®

Published by
Woods Hole Oceanographic Institution
one of the world's leading centers of ocean research

Marine Scientists talking directly to the serious student of the sea

SUBSCRIPTION DEPT. BOX H
1172 COMMONWEALTH AVENUE
BOSTON, MASSACHUSETTS 02134
Please make checks payable to Woods Hole Oceanographic Institution, Woods Hole, Massachusetts 02543.

Please enter my subscription to OCEANUS for one year at $15.00 (four issues).
☐ payment enclosed ☐ bill me

Name.....Address .....

City.....State.....Zip .....

METAMAGICAL THEMAS

The pleasures of Lisp: the chosen language of artificial intelligence

by Douglas R. Hofstadter

In previous columns I have written often about the field of artificial intelligence: the search for ways to program computers so that they might come to behave with such characteristics as flexibility, common sense, insight, creativity, self-awareness and humor. The quest for AI began in earnest more than two decades ago, and since then it has bifurcated many times, so that today it is an active and multifaceted research area. In the U.S. there are perhaps a couple of thousand people professionally involved in AI, and abroad there are a similar number. Although there is among these workers a considerable divergence of opinion concerning the best route to AI, one thing that is nearly unanimous is the choice of programming language. Most AI efforts are carried out in the language called Lisp. The name stands for "list processing."

Why is most AI work done in Lisp? There are many reasons, most of which are somewhat technical, but one of the best is quite simple: Lisp is crisp. Every computer language has arbitrary features, and most languages are in fact overloaded with them. A few, however, such as Lisp and Algol, are built around a kernel that seems as natural as a branch of mathematics. The kernel of Lisp has a crystalline purity that not only appeals to the aesthetic sense but also makes Lisp a language far more flexible than most others. Because of Lisp's beauty and centrality in this important area of modern science, I shall be devoting some columns to presenting some of the basic ideas of the language.

The deep roots of Lisp lie principally in mathematical logic. Mathematical pioneers such as Thoralf Skolem, Kurt Gödel and Alonzo Church contributed seminal ideas to logic in the 1920's and 1930's that were incorporated decades later into Lisp. Computer programming in earnest began in the 1940's, but "higher level" programming languages (of which Lisp is one) came into existence only in the 1950's. The earliest list-processing language was not Lisp but IPL ("information-processing language"), developed in the mid-1950's by Herbert A. Simon, Allen Newell and J. C. Shaw. In the years 1956–58 John McCarthy, drawing on all these earlier sources, came up with the elegant algebraic list-processing language he named Lisp. It caught on quickly with the young crowd around him at the newly formed M.I.T. Artificial Intelligence Project, was implemented on the IBM 704, spread to other AI groups and has stayed around all these years. Many dialects now exist, but all of them share that elegant kernel.

Let us now move on to the way Lisp really works. One of the most appealing features of Lisp is that it is "interactive," as contrasted with most other higher-level languages, which are noninteractive. What this means is the following. When you want to program in Lisp, you sit down at a terminal connected to a computer that has a Lisp system in its memory and you type the word "lisp" (or words to that effect). The next thing you will see on your screen is a "prompt": a characteristic symbol such as an arrow or asterisk. I like to think of the prompt as a greeting spoken by a special "Lisp genie," bowing low and saying to you, "Your wish is my command, and now, what is your next wish?" The genie then waits for you to type something to it. This genie is usually referred to as the Lisp *interpreter,* and it will do anything you want—but you have to take great care in expressing your desires precisely, or you may get some disastrous effects. Shown below is the prompt of Franz Lisp (a particular dialect), the sign that the Franz Lisp genie is ready to do our bidding:

```
—>
```

The genie is asking us for our heart's desire, so let us type in a simple expression,

```
—> (plus 2 2)
```

and then a carriage return. Even non-Lispers can probably anticipate that the Lisp genie will print in return the symbol "4". Then it will also print a fresh prompt, so that the screen will now show:

```
—> (plus 2 2)
4
—>
```

The genie is now ready to carry out our next command or, more politely, our next wish. The carrying out of a wish expressed as a Lisp statement is called the *evaluation* of that statement. The preceding short interchange between human being and computer exemplifies the behavior of the Lisp interpreter: it *reads* a statement, *evaluates* it, *prints* the appropriate value and then signals its readiness to read a new statement. For this reason the central activity of the Lisp interpreter is referred to as the "read-eval-print" loop.

The existence of the Lisp genie (the Lisp interpreter) is what makes Lisp interactive. You get immediate feedback as soon as you have typed a wish—a complete statement—to Lisp. And the way to get a bunch of wishes carried out is to type one, ask the genie to carry it out, then type another, ask the genie again, and so on.

In contrast, in many higher-level computer languages (but by no means all) you must write out an entire program consisting of a vast number of wishes to be carried out in some specified order. What is worse is that later wishes usually depend strongly on the consequences of earlier wishes—and of course you do not get to try them out one by one. The execution of such a program may, needless to say, lead to unexpected results, because so many wishes have to mesh perfectly. If you have made the slightest conceptual error in designing your wish list, then a total foul-up is likely; in fact it is almost inevitable. Running a program of this type is like launching a new space probe untested: you cannot possibly have anticipated all the things that might go wrong, and so all you can do is sit back and watch, hoping it will work. If it fails, you go back and correct the one thing the failure revealed and then try another launch. Such a gawky, indirect, expensive way of programming is in marked contrast to the direct, interactive, one-wish-at-a-time style of Lisp, which allows "incremental" program development and debugging. This is another major reason for the popularity of Lisp.

What kinds of wishes can you type to the Lisp genie for evaluation, and what kinds of things will it print back at you? Well, to begin with, you can type arithmetic expressions expressed in a rather strange way, such as "(times (plus 6 3) (difference 6 3))". The answer is 27, since "(plus 6 3)" evaluates to 9, "(differ-

CIMARRON RESPONDS

You can sense it the moment you first see Cimarron's bold, contemporary road stance. You can feel it the second you slide behind the wheel and into Cimarron's leather-faced, lumbar-supported front bucket seats. And you can experience it every time you take to the open road.

Cimarron responds. See for yourself at your Cadillac dealer's soon. And be sure to ask about the special edition Cimarron D'ORO shown above.

BEST OF ALL... IT'S A CADILLAC.

GM

...AT WORK BY CAR...ION U.S.A.

Exxon introduces offshore oi

Joe Burkhardt helped design a production facility that works on the floor of the sea.

Some of the world's untapped petroleum may be trapped in rocks under water too deep for conventional platforms. That's why Joe Burkhardt and teams of Exxon scientists and engineers have pioneered the development of new technology to take production into deeper and deeper waters.

Breaking new ground beneath the sea

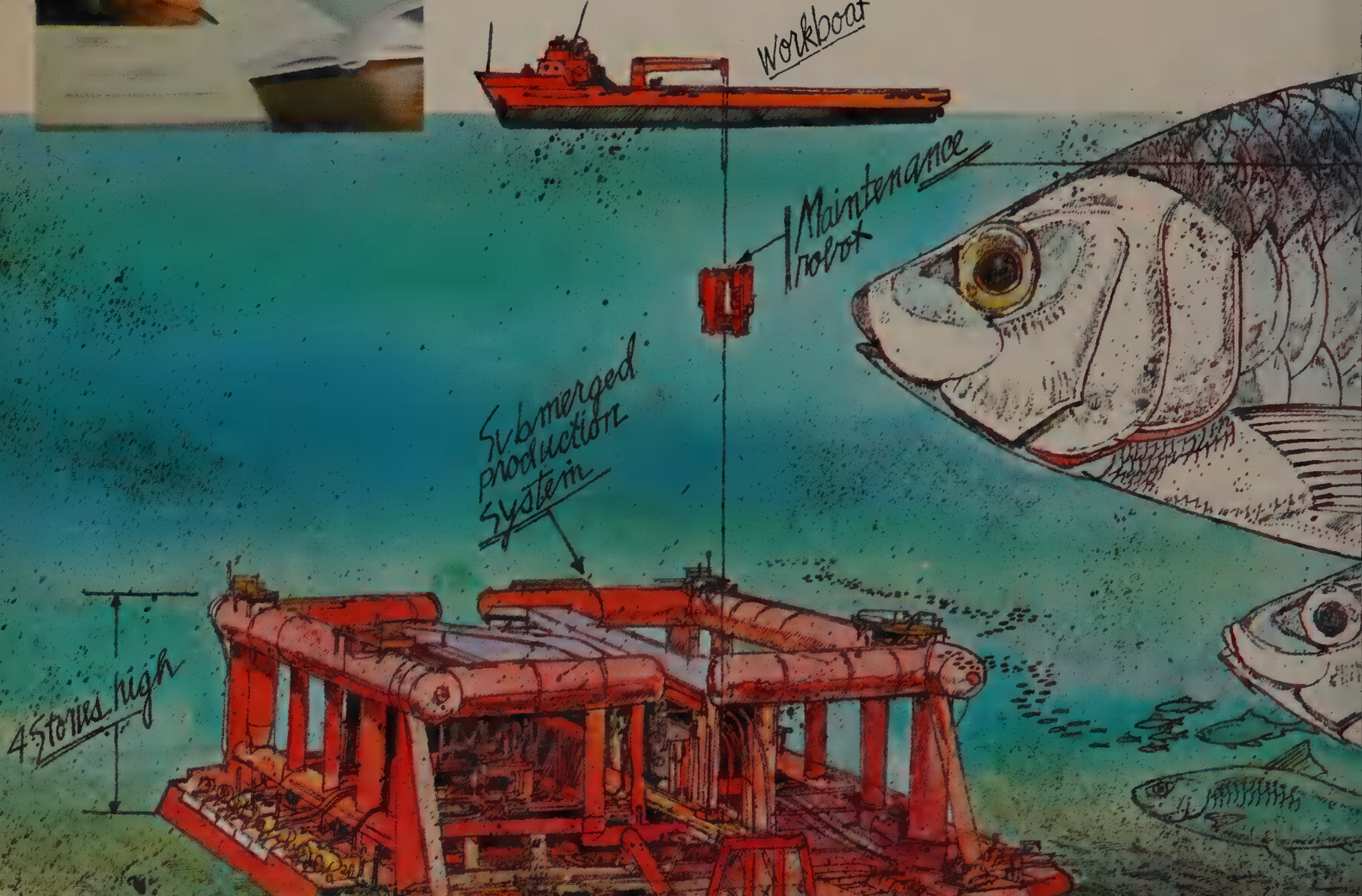
In 1968, Exxon leased tracts off the coast of Southern California, including some under 1800 feet of water. This made it necessary to study ways of producing oil in waters too deep to build platforms. These studies began in a flooded pit near Ventura, California, and culminated in the Gulf of Mexico with the 1974-79 pilot test of Exxon's Submerged Production System—a massive manifold set on the floor of the Gulf.

This prototype SPS controlled thre producing wells drilled through openings in its base and delivered the oil to a nearby platform. Althoug the test was conducted in only 170 feet of water, it proved that the syste could produce oil almost anywhere i the sea where a well could be drille

Now, three years later, the first commercial application of SPS technology is on the floor of the Nort Sea, installed by Exxon and another company.

Subsea maintenance by robot

SPS's have the potential to operate in water depths inaccessible to human divers. So Burkhardt and his



a new step in technology.

olleagues developed sophisticated obotics to perform routine aintenance.

The robot Remote Maintenance ystem (RMS) lowers itself along a uoyed cable from a surface vessel the SPS. It then travels along a ack which gives it access to critical alves and control components. The MS can replace faulty parts, erform pressure tests and carry out arious other maintenance functions. he entire robot operation is bserved on closed-circuit television nd remotely controlled from the rface vessel.

The SPS and the environment

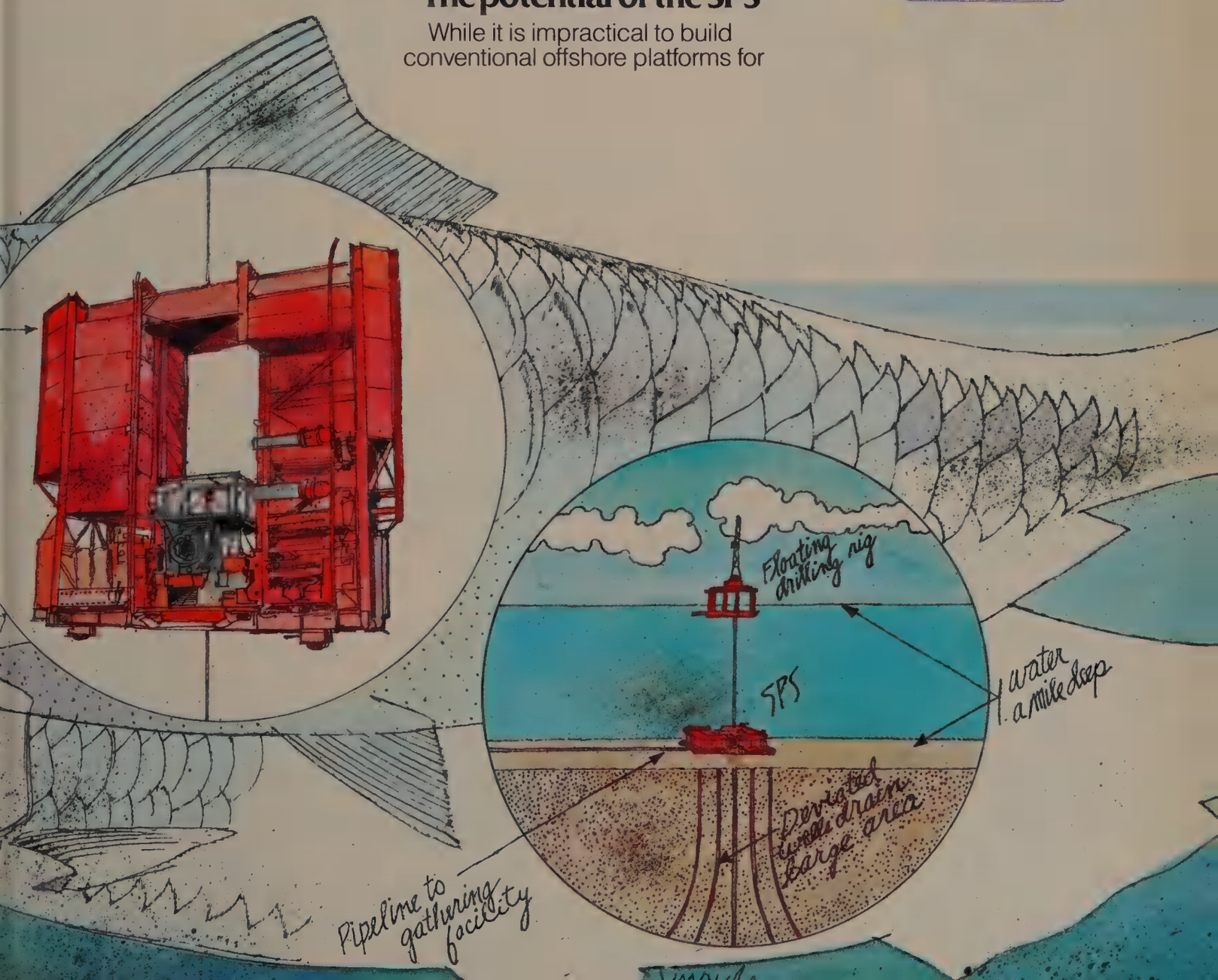
Safeguards incorporated in the SPS reflect Exxon's continuing concern for the environment. A high degree of redundancy is built into many of its subsystems. Fail-safe electrohydraulic control and shut-in systems are used throughout. In the unlikely event that leakage should occur, manifolding on the SPS is covered by inverted "pans" which would collect the leaked oil, trigger an alarm, and stop the leak by closing valves.

The potential of the SPS

While it is impractical to build conventional offshore platforms for water depths much beyond 1,000 feet, SPS technology could be practical to depths of a mile or more. If this full potential is realized, the SPS could be used to open up new areas for oil and gas recovery—areas that cannot be tapped today.

The design and construction of sophisticated deepwater energy systems is just one example of the research and development carried out by Exxon.

EXXON



DISCOVER THE TASTE OF THE RENAISSANCE.

·AMARETTO DI SARONNO·ORIGINALE·1525·

FOR FREE FOOD AND DRINK RECIPE BOOKS, WRITE BOOKLETS, P.O. BOX 99085, LOUISVILLE, KY. 40299

IMPORTED BY ILLVA DISTRIBUTOR CORP., [illegible] ©1985 ENJOY RESPONSIBLY

ence 6 3)" evaluates to 3 and their product is 27. This notation, in which each operation is placed to the left of its operands, was invented by the Polish logician Jan Łukasiewicz before computers existed. Unfortunately for Lukasiewicz his name was too formidable-looking for most speakers of English, and so this type of notation came to be called "Polish notation." Now, here is a simple problem in this notation for you, in which you are to play the part of the Lisp genie:

```
—> (quotient
      (plus 21 13)
      (difference
        23
        (times
          2
          (difference 7 (plus 2 2)))))
```

Perhaps you have noticed that statements of Lisp involve parentheses. A profusion of parentheses is one of the hallmarks of Lisp. It is not uncommon to see an expression that terminates in a dozen parentheses! This makes many people shudder at first, and yet once you get used to their characteristic appearance Lisp expressions become remarkably intuitive, even charming, to the eye, particularly when they are "pretty-printed," which means that a careful indentation scheme is followed that reveals their logical structure (as in the preceding expression).

The heart of Lisp is its manipulable structures. All programs in Lisp work by creating, modifying and destroying structures. Structures come in two types: atomic and composite, or, as they are usually called, *atoms* and *lists*. Thus every Lisp object is either an atom or a list (but not both). The only exception is the special object called "nil", which is both an atom and a list. More about nil below. What are some other typical Lisp atoms? Here are a few:

hydrogen, helium, j-s-bach, 1729, pi, foo, bar, baz, buttons-&-bows

Lists are the flexible data structures of Lisp. A list is pretty much what it sounds like: a collection of some parts in a specific order. The parts of a list are usually called its *elements* or *members*. What can these members be? Well, not surprisingly, lists can have atoms as members. Just as easily, however, lists can contain lists as members, and those lists can in turn contain other lists as members, and so on recursively. Oops! I jumped the gun with that word, but no harm done. You certainly understood what I meant, and it will prepare you for a more technical definition of the term to come later.

A list printed on your screen is recognizable by its parentheses. In Lisp anything bounded by matching parentheses constitutes a list. For instance, "(zonk blee strill (cronk flonk))" is a four-element list whose last element is itself a two-element list. Another short list is "(plus 2 2)", illustrating the fact that Lisp statements themselves are lists. This is important because it means that the Lisp genie, by manipulating lists and atoms, can actually construct new wishes by itself. Hence the object of a wish can be the construction—and subsequent evaluation—of an entirely new wish!

Then there is the empty list: the list with no elements at all. How is this written down? You might think that "()" would work. Indeed, it will work, but there is a second way of indicating the empty list, and that is by writing "nil". The two notations are synonymous, although "nil" is more commonly written than "()". The empty list, nil, is a key concept of Lisp; in the universe of lists it is what zero is in the universe of numbers. To use another metaphor for nil, it is like the earth in which all structures are rooted. To understand what this means, however, you will have to wait a bit longer.

The most commonly exploited feature of an atom is that it has (or can be given) a value. Some atoms have permanent values; others are variables. As you might expect, the value of the atom "1729" is the integer 1729, and this is permanent. (I am distinguishing here between the atom whose "print name" or "pname" is the four-digit string "1729" and the eternal Platonic essence that happens to be the sum of two cubes in two different ways.) The value of nil is also permanent, and it is—nil! One other atom that has itself as its permanent value is the special atom "t".

Apart from t, nil and atoms whose names are numerals, atoms are generally variables, which means that you can assign values to them and later change their values at will. How is it done? Well, to assign the value 4 to the atom "pie" you can type to the Lisp genie "(setq pie 4)". Or you could just as well type "(setq pie (plus 2 2))", or even "(setq pie (plus 1 1 1 1))". In any of these cases, as soon as you type your carriage return, pie's value will become 4, and so it will remain forevermore, or at least until you perform another setq operation on the atom "pie".

Lisp would not be crisp if the only values atoms could have were numbers. Fortunately an atom's value can be set to any kind of Lisp object—any atom or list whatsoever. For instance, we might want to make the value of the atom "pi" be a list such as "(a b c)" or perhaps "(plus 2 2)" instead of the number 4. The value belonging to an atom is often called the atom's *binding*, and one says the atom is *bound* to that value. To assign new bindings to the atoms "pie" and "pi" we again use the "setq" operation. To illustrate, here follows a brief conversation with the genie:

```
—> (setq pie (plus 2 2))
4
—> (setq pi '(plus 2 2))
(plus 2 2)
—>
```

Notice the vast difference between the values assigned to the atoms "pie" and "pi" as a result of these two wishes asked of the Lisp genie, which differ merely in the presence or absence of the small but critical single-quote mark in front of the inner list "(plus 2 2)". In the first wish, since there is no quote mark, that inner "(plus 2 2)" must be *evaluated*. This returns 4, which is assigned to the variable "pie" as its new value. On the other hand, in the second wish, since the quote mark is there, the list "(plus 2 2)" is never executed as a command but is treated merely as an inert lump of Lispstuff, much like meat on a butcher's shelf. It is ever so close to being "alive," yet it is dead. Therefore the value of pi in this second case is the list "(plus 2 2)", a fragment of Lisp code. The following interchange with the genie confirms the values of these atoms:

```
—> pie
4
—> pi
(plus 2 2)
—> (eval pi)
4
—>
```

What is this last step? I wanted to show how you can ask the genie to *evaluate* the value of an expression rather than simply *printing* the value of that expression. Ordinarily the genie automatically performs just one level of evaluation, but by writing "eval" you can get a second stage of evaluation carried out. (And of course by using eval over and over again you can continue as long as you like.) This feature often proves invaluable, but it is a little too advanced to discuss further at this stage.

Every list except nil has at least one element. This first element is called the list's "car". Thus the car of "(eval pi)" is the atom "eval". The cars of the lists "(plus 2 2)", "(setq x 17)", "(eval pi)" and "(car pi)" are all names of operations or, as they are more commonly called in Lisp, *functions*. The car of a list need not be the name of a function; it need not even be an atom. For instance, "((1) (2 2) (3 3 3))" is a perfectly fine list. Its car is the list "(1)", whose car in turn is not a function name but merely a numeral.

If you were to remove a list's car, what would remain? A shorter list. This is called the list's "cdr", a word that is spoken so that it sounds about halfway be-

IF A CAR IS REALLY AN EXTENSION OF ONE'S PERSONALITY, WHAT KIND OF PERSON WOULD DRIVE A SAAB?

Not long ago, a leading car magazine called Saab owners "the lunatic fringe of the American car-buying public."

Yet according to our statistics, the average Saab owner is male, age 38, college-educated, works in a managerial job, and earns over $40,000 a year. He is married and has 1.2 children.

The fact is, both descriptions are accurate.

The fringe.

Some people call this person a driving enthusiast; others call him a car nut.

Whatever you call him, he buys a car for one reason.

Economy? Who cares. Luggage space? Who needs it.

His attitude is if a car doesn't give you goose bumps when you drive it, what's the point of owning it.

For him, even a drive to the supermarket should be exhilarating.

For that, Saab's front-wheel drive and taut suspension give him the cornering ability of a sports car.

And every time Saab's new APC turbocharger kicks in, he feels like he's just engaged warp drive.

Engineering philosophy doesn't interest him. Results do.

Often, he belongs to a car club.

Not the kind with leather jackets and secret handshakes.

But every month or so, they sponsor an event called an Autocross. Much to the dismay of the local townspeople, club members roar their Saabs against the clock through staid suburban parking lots.

Beyond the fringe.

At the other end of the spectrum is the Saab owner who is largely responsible for the respectable statistics that were cited earlier.

He bristles at Saab's cult car reputation. He thinks of car clubs in the same light as motorcycle gangs.

Nonetheless, he does realize that many of Saab's "radical" innovations like turbocharging, front-wheel drive, and aerodynamic design have broader applications than just blowing your neighbor's BMW off the road.

He sees the safety in high performance every time he merges onto a crowded freeway or passes a truck on a two-lane highway.

And, in a Saab APC Turbo, this performance is attained without sacrificing fuel economy. In fact, the APC system actually improves gas mileage.*

He sees the logic of Saab's front-wheel drive and four-wheel disc brakes, especially after the first snowfall. Or the last rainfall.

Even Saab's hatchback design, which some find unconventional, he finds practical, considering that it gives his Saab the carrying capacity of a station wagon.

And not only does his Saab have plenty of room for luggage, it also has plenty of room for people. More, in fact, than many elitist cars.

For those who insist on luxury for luxury's sake, Saab has made one concession. Some turbo models are now equipped with an *Exclusive Appointments Group* that includes leather-upholstered seats and electric sunroofs. (That's really two concessions, isn't it?)

*1983 SAAB PRICE** LIST*

900 3-door	*$10,750*
900 4-door	*$11,050*
900S 3-door	*$13,550*
900S 4-door	*$13,950*
900 Turbo 3-door	*$16,510*
900 Turbo 4-door	*$16,910*

Automatic transmission $370 additional.

Even with leather upholstery and sunroofs you don't have to open manually, Saabs have not replaced Mercedes and BMW as the standard-bearer at the country club.

But for Saab owners, whatever type they may be, the experience of driving a Saab outweighs the lure of status.

It has to.

How else could they get a practical car that drives as well as most wildly impractical cars? A car that appeals to their emotions as well as their intellect?

So what kind of person drives a Saab?

A very satisfied one.

**Saab 900 5-speed APC Turbo: (21) EPA estimated mpg, 34 estimated highway mpg. Use estimated mpg for comparison only. Mileage varies with speed, trip length and weather. Actual highway mileage will probably be less. **Manufacturer's suggested retail prices. Not including taxes, license, freight, dealer charges or options.*

SAAB
SAAB
The most intelligent car ever built.

PSST!
WANT AN UNLISTED TELEPHONE NUMBER?

We can give you one with the press of a button.
That's right. The press of a button.

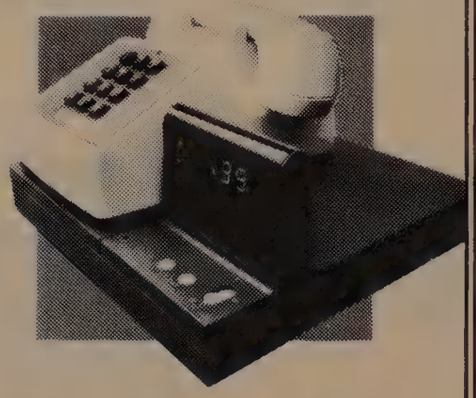
No, we're not Ma Bell. We're International Mobile Machines Corporation. And with the press of a button, we make it possible for you to create, not one, but as many as 16 separate, unlisted, private telephone numbers...whenever you want them, as often as you want them.

How? With PriveCode®– the Telephone Access Control Terminal™...an affordable, convenient alternative to Ma Bell's unlisted number system.

Here's how it works:
Just plug PriveCode into any modular jack in your home or apartment, turn it on, and you're in complete control over who can reach you. You simply assign a 3 digit Access Code to people you want to hear from. Whenever anyone dials your regular listed telephone number, PriveCode intercepts their call and **stops all your phones from ringing.** PriveCode then asks the caller to enter their Personal Access Code. Only after a valid code has been entered will PriveCode signal you with a pleasing electronic sound, and display the caller's personal code, so you know who is calling before you pick up your phone. PriveCode even tells you who called while you were out. Callers without valid Access Codes never get through. Since PriveCode handles up to 16 different codes at a time it's just like having 16 private lines...on your present Ma Bell line. So say "Goodbye" to wrong numbers and heavy breathers...say "Hello"...only when you want to.

To create the privacy, and the security of an unlisted telephone number with the press of a button, say "Hello" to PriveCode.

PriveCode is available in fine department stores, electronic specialty shops and phone stores.



PRIVECODE®
The Telephone Access Control Terminal™

See a PriveCode demonstration today, call toll-free for the PriveCode dealer nearest you.

800-523-0103, Ext. 110
(In PA call 215-569-1300, ext. 110)

For further information write
International Mobile Machines Corporation
100 North 20th Street, Department No. 111
Philadelphia, Pennsylvania 19103

New Management Tool!

It's called the **Executive ScanCard™ System,** and it serves as a portable "control center" for monitoring all your projects on a daily basis, so you can double, even triple, the number of projects you can keep moving simultaneously.

Easy to use
Projects are recorded on ScanCard™ project cards that can be quickly scanned to jog your memory on what's to be done next.

Expandable
Initially holds 64 projects. More panels (40 cards each) and ScanCards can be ordered as needed.

The "Executive Look"
In handsome, rugged leather look material that will last for years. Available in black, brown, tan, burgundy or natural suede.

National Introductory Price
$39.95
Add $4 shipping and handling
Ohio residents add sales tax.

Your name printed here.

Important extras
Includes pen holder and space for your tablet. Plus 500 ScanCards and illustrated instruction booklet.

FREE
Handy **Pocket Idea™** file for carrying small supply of ScanCards. Jot your ideas on cards and merge into system later.

Satisfaction Guaranteed
If for any reason you are not satisfied, you may return the system within 30 days for a full refund.

To order phone
800-848-2618
IN OHIO CALL 800-282-2630

or mail check to
Executive ScanCard™ Systems Dept. 748
8480 Busch Blvd. Suite 200
Columbus, Ohio 43229

We honor American Express, MasterCard, Visa and Diners Club.

International Orders — Shipping & Handling Charges:
Canada — $7.50, Mexico — $12.00, All other International — $18.00.
International payments in U.S. Dollars.

PATENTS PENDING©COPYRIGHT 1981 EXECUTIVE SCANCARD™ SYSTEMS

tween "kidder" and "could 'er." (The words "car" and "cdr" are relics from the first implementation of Lisp on the IBM 704. The letters in "car" and "cdr" stand respectively for "contents of the address part of register" and "contents of the decrement part of register." They refer to specific hardware features of that machine, now long since irrelevant.) The cdr of "(a b c)" is the list "(b c)", whose cdr is "(c)", whose cdr is nil. And nil has no cdr, just as it has no car. Attempting to take the car or cdr of nil causes (or should cause) the Lisp genie to cough out an error message, just as attempting to divide by zero should evoke an error message.

Here is a little table showing the cars and cdrs of a few lists, just to make sure the notion is unambiguous:

list	*car*	*cdr*
((a) b (c))	(a)	(b (c))
(plus 2 2)	plus	(2 2)
((car x) (car y))	(car x)	((car y))
(nil nil nil nil)	nil	(nil nil nil)
(nil)	nil	nil
nil	ERROR	ERROR

Just as car and cdr are called *functions,* so the things they operate on are called their *arguments.* Hence in the command "(plus pie 2)", "plus" is the function name and the arguments are the atoms "pie" and "2". In evaluating this command (and most other commands) the genie figures out the values of the arguments and then applies the function to those values. Since the value of the atom "pie" is 4 and the value of the atom "2" is 2, the genie returns the atom "6".

Suppose you have a list and you would like to see a list just like it, only one element longer. For instance, suppose the value of the atom "x" is "(cake cookie)" and you would like to create a new list called "y" just like x, except with an extra atom—say "pie"—at the front. You can then use the function called "cons" (short for "construct"), whose effect is to make a new list out of an old list and a suggested car. Here is a transcript of such a process:

```
—> (setq x '(cake cookie))
(cake cookie)
—> (setq y (cons 'pie x))
(pie cake cookie)
—> x
(cake cookie)
```

Two things are worth noticing here. I asked for the value of x to be printed out after the cons operation, so that you could see that x itself was not changed by that operation. Cons created a new list and made that list be the value of y, but it left x entirely alone. The other noteworthy fact is that I used that quote mark again, in front of the atom

EXTINCTION. IT'S FOREVER.

One quarter of all species of animals and plants on Earth may disappear in the next 30 years because of man's destruction of their habitat. The rate of extinctions is increasing enormously as forests are destroyed and other wild areas are lost. Organisms that evolved over hundreds of millions of years will be gone forever. The complex interdependence of all creatures, from the largest mammals to the smallest plants, is being shattered. It is a crisis with profound implications for the survival of all life. Unfortunately, little is being done to save our planet's natural heritage. Here are some warnings by leading scientists:

"The worst thing that can happen—will happen—in the 1980's is not energy depletion, economic collapse, limited nuclear war, or conquest by a totalitarian government. As terrible as these catastrophes would be for us, they can be repaired within a few generations. The one process ongoing in the 1980's that will take millions of years to correct is the loss of genetic and species diversity by the destruction of natural habitats. This is the folly our descendants are least likely to forgive us."

DR. EDWARD O. WILSON
Baird Professor of Science • Harvard University

"We are encroaching on nature, in the U.S. and around the world, at an unprecedented rate. A large proportion of the chemicals in use in our present-day civilization were 'invented' by nature, not by the chemist in the laboratory. An estimated 40% of all drug prescriptions in the U.S. contain as their chief ingredients compounds derived from plants. There is no end to the potential for discovery in nature, because we have only begun the chemical exploration of nature. Tragically, we are burning our library of priceless genetic treasures with our reckless destruction of species."

DR. THOMAS EISNER
Schurman Professor of Biology • Cornell University

"Few problems are less recognized but more important than the accelerating disappearance of Earth's biological resources. In pushing other species to extinction, humanity is busily sawing off the limb on which it is perched."

DR. PAUL EHRLICH
Bing Professor of Population Studies • Stanford University

"The extermination of a quarter of the plant species on Earth during the coming few decades could lose forever the medicines and food sources we need to cure the diseases and hunger that plague mankind."

DR. PETER RAVEN
Director • Missouri Botanical Garden

The Ark is sinking. We need the help of every concerned citizen to conserve the diversity of life on Earth. For information about how you can help, please write to us at:

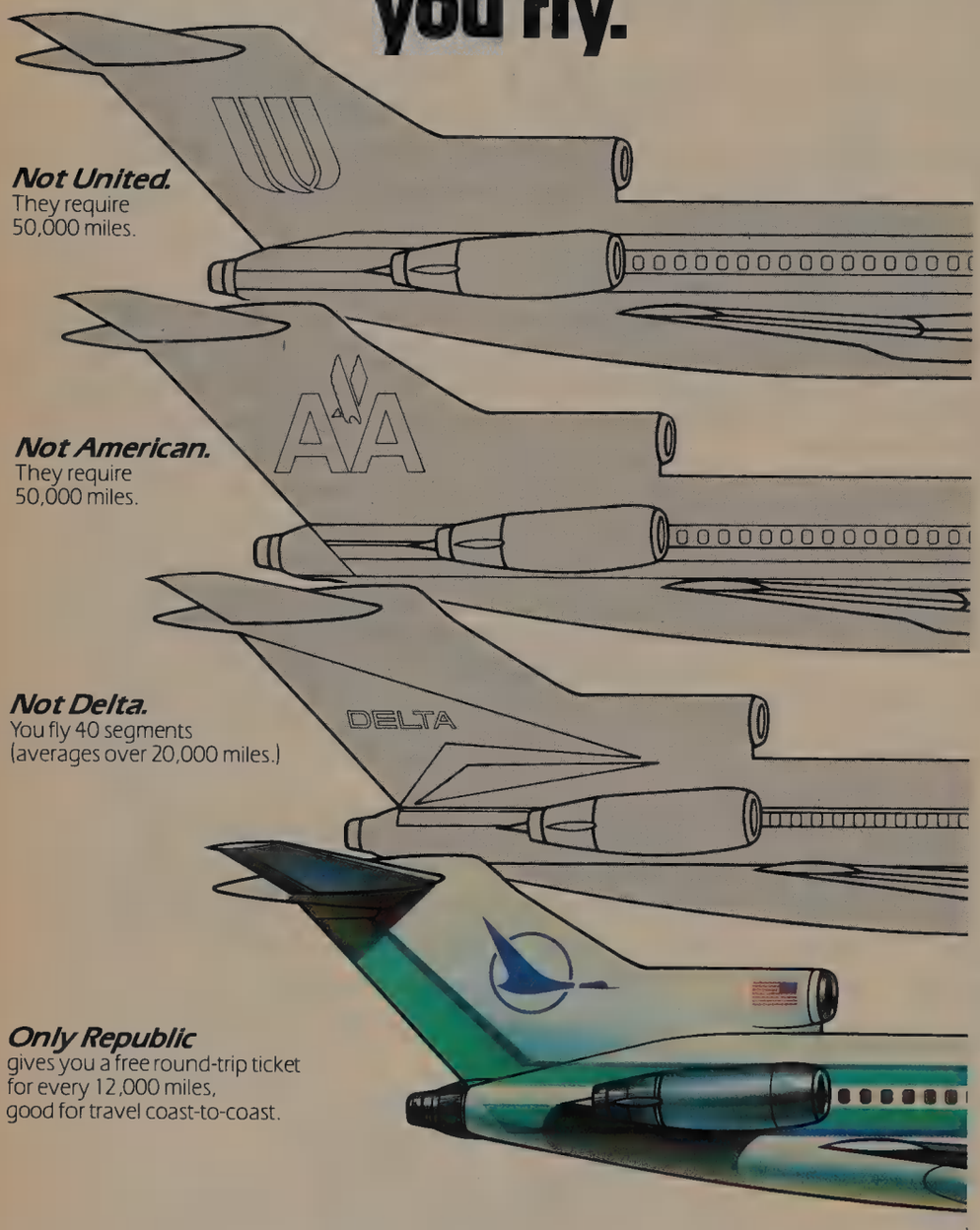
Save Endangered Species

P.O. Box 50771 • Washington, D.C. 20004

Sponsored by:

- American Association of Zoological Parks and Aquariums
- American Cetacean Society
- American Humane Association
- American Institute of Biological Sciences
- Animal Protection Institute of America
- California Native Plant Society
- Center for Action on Endangered Species
- Center for Environmental Education
- Defenders of Wildlife
- Desert Fishes Council
- Environmental Defense Fund
- Florida Audubon Society
- Friends of the Earth
- Friends of the Sea Otter
- Fund for Animals
- Garden Club of America
- Greenpeace U.S.A.
- Humane Society of the United States
- International Council for Bird Preservation
- Massachusetts Audubon Society
- National Audubon Society
- National Parks and Conservation Association
- National Shell Fisheries Association
- Natural Resources Defense Council
- Ocean Research and Education Society
- Sierra Club
- Society for Animal Protective Legislation
- Wilderness Society
- World Wildlife Fund - U.S.

Nobody
but Republic gives you free travel for every 12,000 miles you fly.



Not United.
They require
50,000 miles.

Not American.
They require
50,000 miles.

Not Delta.
You fly 40 segments
(averages over 20,000 miles.)

Only Republic
gives you a free round-trip ticket
for every 12,000 miles,
good for travel coast-to-coast.

**The new Republic Frequent Flyer™ Program:
enroll now and begin earning your free round-trip travel.**

Name _____ Company _____

Phone: Home ( ) _____ Business ( ) _____

Do you prefer to receive your Frequent Flyer mail at: ☐ Home ☐ Business

Address _____

City _____ State _____ Zip _____

Mail to Republic Airlines Frequent Flyer Program, Dept. SA , P.O. Box 20452, Atlanta, GA 30325.
We'll forward your personal account number and full information.

NOBODY SERVES OUR REPUBLIC LIKE REPUBLIC™ AIRLINES

"pie". What if I had not used it? This is what would have happened:

```
—> (setq z (cons pie x))
(4 cake cookie)
```

Remember, after all, that the atom "pie" still has the value 4, and whenever the genie sees an unquoted atom inside a wish, it will always use the value belonging to that atom rather than the atom's name. (Always? Well, almost always. I shall explain in a moment. In the meantime look for an exception; you have already encountered it.)

Now here are a few exercises for you, some of them a bit tricky. Watch out for the quote marks! One last thing: I use the function "reverse", which produces a list just like its argument, only with its elements in reverse order. For instance, the genie, on being told "(reverse '((a b) (c d e)))", will write "((c d e) (a b))". The genie's lines in this dialogue are given afterward.

```
—> (setq w (cons pie '(cdr z)))
—> (setq v (cons 'pie (cdr z)))
—> (setq u (reverse v))
—> (cdr (cdr u))
—> (car (cdr u))
—> (cons (car (cdr u)) u)
—> u
—> (reverse
       '(cons (car u) (reverse (cdr u))))
—> (reverse
       (cons (car u) (reverse (cdr u))))
—> u
—> (cons 'cookie
          (cons 'cake
                (cons 'pie nil)))
```

The answers (as printed by the genie) are as follows:

```
(4 cdr z)
(pie cake cookie)
(cookie cake pie)
(pie)
cake
(cake cookie cake pie)
(cookie cake pie)
((reverse (cdr u)) (car u) cons)
(cake pie cookie)
(cookie cake pie)
(cookie cake pie)
```

The last example, featuring the repeated use of cons, is often called, in Lisp slang, "consing up a list." You begin with nil and then do repeated cons operations. It is analogous to building a positive integer by starting at zero and then performing the successor operation over and over again. However, whereas at any stage in the latter process there is a unique way of doing the successor operation, given any list there are infinitely many different items you might cons onto it, thus giving rise to a vast branching tree of lists instead of the unbranching number line. It is because of this im-

age of a tree growing out of the ground of nil and containing all possible lists that I likened nil to the earth in which all structures are rooted.

As I mentioned above, the genie does not *always* replace (unquoted) atoms by their values. There are cases where a function treats its arguments, although unquoted, as if they were quoted. Did you go back and find such a case? It is not hard. The answer is the function "setq". In particular, in a setq command the first atom is taken straight, not evaluated. As a matter of fact, the "q" in "setq" stands for "quote," meaning that the first argument is treated as if it were quoted. Things can get tricky when you learn about "set", a function similar to setq except that it *does* evaluate its first argument. Hence if the value of the atom "x" is the atom "k", then entering "(set x 7)" will not do anything to x (its value will remain the atom "k") but the value of the atom "k" will now become 7. Watch closely:

```
—> (setq a 'b)
—> (setq b 'c)
—> (setq c 'a)
—> (set a c)
—> (set c b)
```

Now tell me: What are the values of the atoms "a," "b" and "c"? Here comes the answer, so try not to peek. They are respectively "a", "a" and "a". This may be confusing. You may be reassured to know that in Lisp, "set" is not very commonly used and such confusions do not arise often.

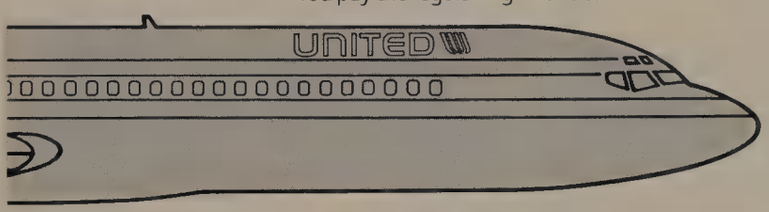
Psychologically, one of the great powers of programming is the ability to define new compound operations in terms of old ones, and to do this over and over again, thus building up a vast repertory of ever more complex operations. It is reminiscent of the early stages of biological evolution, when ever more complex molecules evolved out of less complex ones in an ever upward helix of complexity and creativity. It is also reminiscent of the Industrial Revolution, in which people used simple early machines to help them build more complex machines, then used those in turn to build even more complex machines, again in an ever upward helix of complexity and creativity. At each stage, whether in evolution or Revolution, the products get more flexible and more intricate, more "intelligent" and yet more vulnerable to subtle "bugs" or breakdowns.

So it is with programming in Lisp, except that here the "molecules" or "machines" are Lisp functions defined in terms of previously known Lisp functions. Suppose, for instance, you want to have a function that will always return the last element of a list, just as "car" always returns the first element of a

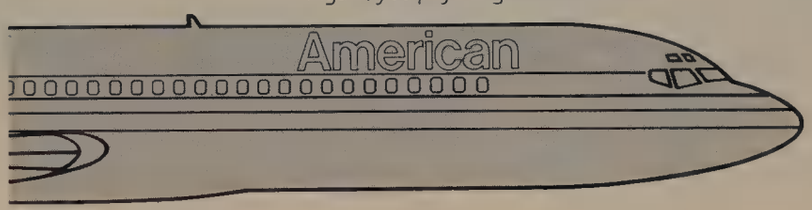
Nobody
but Republic gives you first class comfort for just $10 over Coach.

(No matter how far you're flying.)

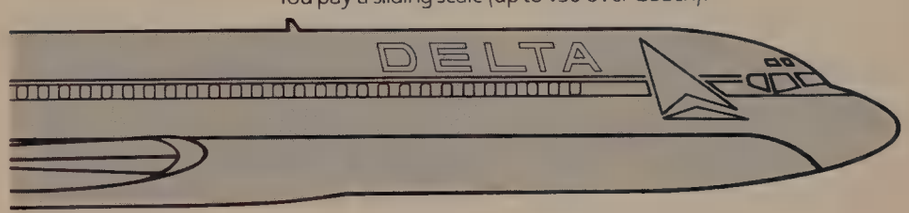
Not United.
You pay the regular high front-cabin fare.



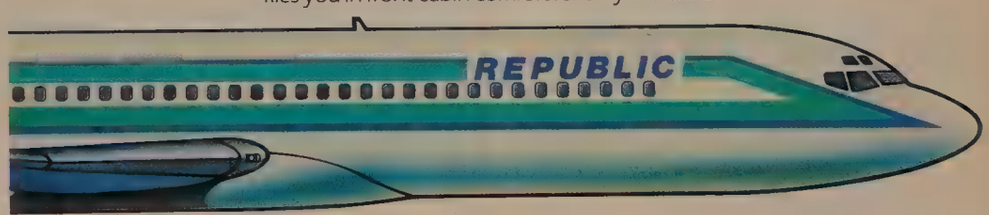
Not American.
Again, you pay a high front-cabin fare.



Not Delta.
You pay a sliding scale (up to $30 over Coach).



Only Republic
flies you in front-cabin comfort for only $10 extra.



Republic Business Coach:™ the fare is as comfortable as the flight.

For a flat $10 more than our one-way nondiscounted Coach fare, you can enjoy all the amenities of Republic Business Coach. The privacy of a separate up-front cabin. A comfortable two-by-two seat. Complimentary beverages. And priority baggage handling. A great place to prepare for the business day. A great place to relax afterward. Available on most flights. Call a travel agent, your corporate travel planner or Republic.

NOBODY SERVES OUR REPUBLIC LIKE REPUBLIC™ AIRLINES

THE CAR THAT TOOK TEN YEARS TO BUILD.

In the past ten years alone, no fewer than 30 "exciting" new cars have been introduced in the U.S., hyped, sold and then quietly discontinued.

Remember the once-famous Monza, Versailles, Pacer, Dasher, Fox, Astre, Monarch, Sunbird, Bobcat, Starfire, Aspen, Volare? Where, oh where, are they now?

In that same ten year period Volvo has been developing a single, remarkable new automobile. The 760 GLE.

As excited as we were about this car, we weren't about to introduce it until it was as close to perfect as we could make it.

So prototypes were driven a total of two million miles through some of the most gruelling conditions on earth. They hauled trailers over the Alps, plowed through snowbanks near the Arctic Circle, pounded through potholes in New York City and ate dust in Australia. They endured five searing summers in the Nevada desert, frequently being driven 24 hours a day.

All this testing led to hundreds of design improvements. And after we'd made those improvements, we improved on the improvements.

What emerged from all this is a car that's astonishing indeed.

The 760 GLE is more aerodynamically efficient than a Porsche 928, yet roomy enough for five adults.

It has traveled through a high-speed slalom course faster than a Jaguar or Audi.

It has a climate control system so advanced it can change the interior air five times a minute while the car is standing still.

The Volvo 760 GLE.

It took a long time to get here. But we wanted to make sure that, unlike so many cars, once it got here it stayed.

THE VOLVO 760 GLE

The closest thing yet to a perfect car.

© 1983 VOLVO OF AMERICA CORPORATION.

list. Lisp does not come equipped with such a function, but you can easily create one. Do you see how? To get the last element of a list called "lyst", you simply reverse lyst and then take the car of that: "(car (reverse lyst))". To dub this operation "rac" ("car" backward) we use the "def" function, as follows:

```
—> (def rac (lambda (lyst)
        (car (reverse lyst))))
```

Using def in this way creates a *function definition*. In it the word "lambda" followed by "(lyst)" indicates that the function we are defining has only one parameter, or dummy variable, to be called "lyst". (It could have been called anything; I just happen to like the atom "lyst".) In general the list of parameters must immediately follow the word "lambda". After this "def wish" has been carried out the rac function is as well understood by the genie as the car function is. Thus "(rac '(your brains))" will yield the atom "brains". And we can use rac itself in definitions of still further functions. The whole thing snowballs rather miraculously, and you can quickly become overwhelmed by the power you wield.

Here is a simple example. Suppose you have a situation where you know you are going to run into many big long lists and you know it will often be useful to form, for each such long list, a short list that includes just its car and rac. We can define a one-parameter function to do it for you:

```
(def readers-digest-condensed-version
   (lambda (biglonglist)
    (cons (car biglonglist)
          (cons (rac biglonglist) nil))))
```

Hence if we apply the function "readers-digest-condensed-version" to the entire text of James Joyce's *Finnegans Wake* (treating it as a big long list of words), we shall obtain the shorter list "(river-run the)". Unfortunately reapplying the condensation function to this new list will not simplify it any further.

It would be nice as well as useful if we could create an inverse operation to readers-digest-condensed-version called "rejoyce" that, given any two words, would create a James Joyce novel beginning and ending with them respectively. Thus execution of the Lisp statement "(rejoyce 'Stately 'Yes)" would result in the Lisp genie's generating from scratch the entire novel *Ulysses*. Writing this function is left as an exercise for the reader.

One goal that has seemed to some people to be both desirable and feasible with Lisp and related programming languages is (1) to make every single statement return a *value* and (2) to have it be through this returned value and *only* through it that the statement has any effect. The idea of (1) is that values are handed "upward" from the innermost function calls to the outermost ones, until the full statement's value is returned to you. The idea of (2) is that during all these calls no atom has its value changed at all (unless the atom is a dummy variable). In all dialects of Lisp that are known to me (1) is true but (2) is not necessarily true.

Thus if x is bound to "(a b c d e)" and you say "(car (cdr (reverse x)))", the first thing that happens is that the value of "(reverse x)" is calculated; then this value is handed "up" to the cdr function, which calculates the cdr of that list; finally this shorter list is handed to the car function, which extracts one element, the atom "d", and returns it. In the meantime the atom "x" has suffered no damage; it is still bound to "(a b c d e)".

It might seem that an expression such as "(reverse x)" would change the value of x by reversing it, just as carrying out the oral command "Turn your sweater inside out" will affect the sweater. Actually carrying out the wish "(reverse x)" no more changes the value of x than carrying out the wish "(plus 2 2)" changes the value of 2. Instead executing "(reverse x)" causes a *new*, unnamed list to come into being, just like x, only reversed. And that list is the value of the statement; it is what the statement returns. The value of x itself, however, is untouched. Similarly, evaluating "(cons 5 pi)" will not change the value of the list named "pi" in the slightest; it merely *returns* a new list with 5 as its car and whatever pi's value is as its cdr.

Such behavior is to be contrasted with that of functions that leave "side effects" in their wake. Such side effects are usually in the form of changed variable bindings, although there are other possibilities, such as causing input or output to take place. A typical "harmful" command is a setq, and proponents of the "applicative" school of programming—the school that says you should never have any side effects whatever—are profoundly disturbed by the mere mention of setq. For them all results must come about purely by the way functions compute their values and hand them along to other functions.

The only bindings the advocates of the applicative style approve of are transitory "lambda bindings": those that arise when a function is applied to its arguments. Whenever any function is called, that function's dummy variables temporarily assume "lambda bindings." These bindings are just like those resulting from a setq except that they are fleeting. That is, the moment the function is finished computing they go away, vanishing without a trace. For example, in the computation of "(rac '(a b c))" the lambda binding of the dummy variable "lyst" is to the list "(a b c)", but as soon as the answer "c" is passed along to the person or function that requested the rac, the value of the atom "lyst" used in getting that answer is totally forgotten. If you ask for the value of lyst, the Lisp interpreter will tell you that the atom "lyst" is an "unbound atom." Applicative programmers much prefer lambda bindings to ordinary setq bindings.

I personally am not a fanatic about avoiding setq's and other functions that cause side effects. Although I find the applicative style elegant, I think it impractical for the construction of large AI programs. Therefore I shall not advocate the applicative style here, although I shall adhere to it when possible. Strictly speaking, in applicative programming you cannot even define new functions, since a "def" statement causes a permanent change to take place in the genie's memory, namely the permanent storage in memory of the function definition. Thus the ideal applicative approach would have function definitions, like variable bindings, being created only temporarily and then discarded the moment after they had been used. This is extreme "applicativism."

For your edification here are a few more simple function definitions:

```
—> (def rdc
        (lambda (lyst)
         (reverse (cdr (reverse lyst)))))
—> (def snoc
        (lambda (x lyst)
         (reverse
          (cons x (reverse lyst)))))
—> (def twice (lambda (n) (plus n n)))
```

The functions rdc and snoc are analogous to cdr and cons, only backward. Hence the rdc of "(a b c d e)" is "(a b c d)", and if you type "(snoc 5 '(1 2 3 4))", you will get as your answer "(1 2 3 4 5)".

All of this is mildly interesting so far, but if you want to see the genie do anything *truly* surprising, you have to allow it to make some decisions based on things that happen along the way. These are sometimes called *conditional* wishes. A typical example would be the following:

```
—> (cond ((eq x 1) 'land)
         ((eq x 2) 'sea))
```

The value returned by this statement will be the atom "land" if x has the value 1, and the atom "sea" if x has the value 2. Otherwise the value returned will be nil (for instance if x is 5). The atom "eq" (pronounced "eek") is the name of a common Lisp function that returns the atom "t" (standing for "true") if its two arguments have the same value, and nil (for "no" or "false") if they do not.

A cond statement is a list whose car is the function name "cond", followed by any number of *cond clauses*, each of

A MAGAZINE FOR YOUR PIANO

Keyboard Classics Magazine *takes up where* ***The Etude*** *magazine left off. Filled with musical treasures, fancies and delights to be played on your piano, it is truly the one magazine you will enjoy again and again.*

Available only by subscription.
Write To: Keyboard Classics,
Dept. A1-05-23, 352 Evelyn Street,
Paramus, NJ 07652.
Enclose $9.97 for a full year's subscription (6 issues).
Completely refundable if you are not delighted with your first issue.

SPOTTING SCOPES

Precision Optics From
CELESTRON®

This 90mm (diameter) Spotting Scope is also a 1000mm mirror lens telephoto! We offer superb quality Spotting Scopes, Binoculars, Telephotos and Astronomical Telescopes in all price ranges. For **FREE** product price and ordering information send coupon below (or send $1.00 for a detailed catalog/selection guide).

CALL TOLL FREE 1-800-421-1525

CELESTRON INTERNATIONAL
P.O. Box 3578-SA, 2835 Columbia St.
Torrance, CA 90503

Name

Address

City State Zip

which is a two-element list. The first element of each clause is its "condition," the second element is its "result." The clauses' conditions are checked out by the Lisp genie one by one, in order; as soon as the genie finds a clause whose condition is "true" (meaning that the condition returns anything other than nil) it begins calculating that clause's *result,* whose value gets returned as the value of the entire cond statement. None of the further clauses is even so much as glanced at! This may sound more complex than it ought to. The real idea is no more complex than saying that the genie looks for the first condition that is satisfied, then returns the corresponding result.

Often one wants to have a catchall clause at the end whose condition is *sure* to be satisfied, so that if all other conditions fail, at least this one will be true and the accompanying result, rather than nil, will be returned. It is as easy as pie to make a condition whose value is non-nil; just choose it to be "t", for instance, as in the following:

```
—> (cond ((eq x 1) 'land)
         ((eq x 2) 'sea)
         (t 'air))
```

Depending on what the value of x is, we shall get either "land", "sea" or "air" as the value of this cond, but we shall never get nil. Here are a few sample cond statements for you to play genie to:

```
—> (cond ((eq (eval pi) pie)
          (eval (snoc pie pi)))
         (t (eval (snoc (rac pi) pi))))
—> (cond ((eq 2 2) 2) ((eq 3 3) 3))
—> (cond (nil 'no-no-no)
         ((eq '(car nil) '(cdr nil))
          'hmmm)
         (t 'yes-yes-yes))
```

The answers are 8, 2 and "yes-yes-yes". Did you notice that "(car nil)" and (cdr nil)" were quoted?

I shall close this part of the column by displaying a family of function definitions so obvious in their pattern that you would think the Lisp genie would simply "get the hang of it" after seeing the first few. Unfortunately Lisp genies are frustratingly dense (or at least they play at being dense), and they will not jump to any conclusion unless it has been spelled out explicitly for them. Look first at the family:

```
—> (def square
       (lambda (k) (times k k)))
—> (def cube
       (lambda (k) (times k (square k))))
—> (def 4th-power
       (lambda (k) (times k (cube k))))
—> (def 5th-power
       (lambda (k)
        (times k (4th-power k))))
—> (def 6th-power
       (lambda (k)
        (times k (5th-power k))))
:
:
:
```

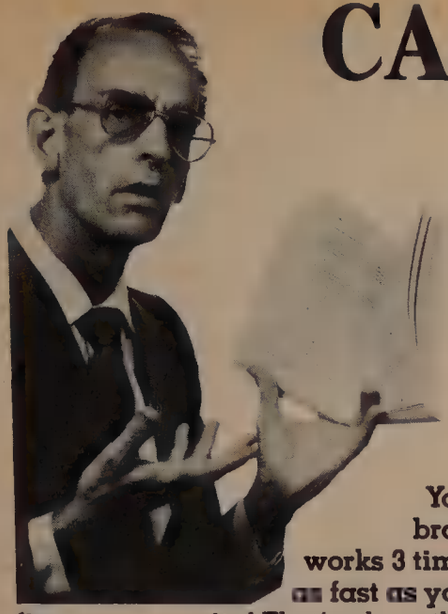
My question for you is this: Can you invent a definition for a *two*-parameter function that subsumes this entire infinite family in one fell swoop? (Please do not write to give me your thoughts; I know the answer and will provide it next time.) More concretely, the question is: How would one go about defining a two-parameter function called "power" such that, for instance, "(power 9 3)" yields 729 on being evaluated and "(power 7 4)" yields 2401? I have given you in this column all the necessary tools to do it, provided you exercise some ingenuity.

I thought I would end with a news-break about a freshly discovered beast: the Glazunkian porpuquine, so called because it is found only on the island of Glazunkia (claimed by Upper Bitbo, although it is just off the coast of Burronymede). And what is a porpuquine, you ask? Why, it is a strange breed of porcupine whose quills—of which, for some reason, there are always exactly nine (in Outer Glazunkia) or seven (in Inner Glazunkia)—are smaller porpuquines. Oho! This would certainly seem to be an infinite regress! But no. It is just that I forgot to mention that there is a smallest size of porpuquine: the zero-inch type, which, amazingly enough, is totally bald of quills. Thus quite luckily (or perhaps unluckily, depending on your point of view) that puts a stop to the threatened infinite regress.

Students of zoology might be interested to learn that the quills on five-inch porpuquines are always four-inch porpuquines, and so on down the line. And students of anthropology might be equally intrigued to know that the residents of Glazunkia (both Outer and Inner) utilize the nose (yes, the nose) of the zero-inch porpuquine as a unit of barter—an odd thing to our minds, but then who are you and I to question the ancient ways of the Outer and Inner Glazunkians? Hence since a largish porpuquine—say a three- or a four-incher—has many, many such tiny noses, it is a most valuable commodity. The value of a porpuquine is sometimes referred to as its "buying power," or just "power" for short. For instance, a two-incher found in Inner Glazunkia is almost twice as powerful as a two-incher found in Outer Glazunkia. Or did I get it backward?

Anyway, why am I telling you all this? Oh, I just thought you would like to hear about it. Anyway, who knows? You just might wind up visiting Glazunkia (Inner or Outer) one of these fine days. And then all of this could come in mighty handy.

CALCULATOR MATH?



Your brain works 3 times as fast as your fingers can write! That's where the Calculator comes in — you can go three times as fast with your machine doing the dirty work.

It's easier and more interesting. And you'll understand it better when it's fun. (If you haven't yet tried out a modern scientific-statistical calculator, you're in for a pleasant surprise — see below!)

WHY MATH? To get a promotion or a new job. Just to hang on to that job you've got. To quicken your mind. To understand this world you live in. To impress people. Just for some fun you'll feel good about. (Read those letters below to see how it has worked for others.)

Do it at home... You can start right away and go as fast as you like! Put it aside while you're busy with something else — in short, *set your own pace!*

Lots of us like to avoid being compared with the other members of a class — we like to set our own standards. Too, we like to save the time and expense of going to a school or seminar. (Best of all, it's a lot cheaper to study at home where the rent is already paid — *read on!*)

Let's be honest. You wouldn't expect to learn to play baseball or chess by watching an expert do it. Math is the same — you'll need to do some work. BUT — with my calculator method, you'll see it's a whole new ball game!

THIS METHOD IS FRIENDLY! You can use it on your own. You follow interesting examples step-by-step with calculator keystrokes — instead of algebra.

Let me introduce myself — I'm Dr. George McCarty and I teach at the University of California. Tens of thousands of people like yourself have already used my guidebook to teach themselves calculus. Recently I have teamed up with another professor. We put those same tested and proven techniques to work for you in a new *Statistics Guidebook.*

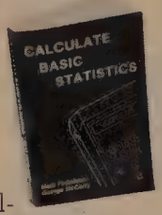
Even if you've had trouble with math, this method will work for you. Choose the one you want to learn, calculus or statistics (or both) and start *today!*

As pennywise Ben Franklin said, "An investment in knowledge pays the best dividends."

CHOOSE STATISTICS

It's used everywhere — to set insurance rates and to assess a drug's effectiveness, to detect a crooked roulette wheel and to measure lightbulb lifetimes, to audit the books and to find your position at sea.

People have always known that their experience was not a perfect guide to the future. Statistics shows us that uncertainty is not only inevitable but it itself is measurable.

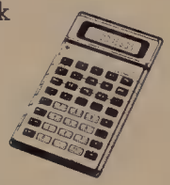


You'll find out about this in my guidebook. You'll read informal explanations with detailed calculations for topics like "confidence intervals" and "contingency tables."

YOU'LL BE DOING Inference and Hypothesis Testing, the Chi-Square Distribution, ANOVA, Regression and Correlation, Non-Parametric Methods — they're all here.

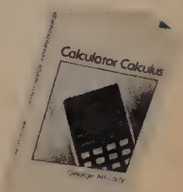
SUPPOSE YOU HAVE ALREADY STUDIED STATISTICS. You'll use Quick-Reference Guides to go straight to a topic you want to refresh yourself on — then get *right-now* answers in your own field, business, biology, science, psychology, education.

The Guidebook is guaranteed to work with any calculator or computer. If you would like to start out with a fancy new machine, order my Statistics Kit which includes the guidebook and a TI-35 calculator which finds means and standard deviations at the touch of a button. It has an LCD display, 1000 hour batteries — the works.



CHOOSE CALCULUS

The math of science and engineering — satellite orbits and compound interest, bacterial growth and radioactive decay, marginal profits and reaction rates.



You'll find out about these concepts when you use the Calculus Kit. This Kit contains my guidebook and a TI-35 calculator which finds log and trig values at the touch of a button. (It's modern, too, with LCD display and 1000 hour batteries.)

In the guidebook you'll learn both differential calculus and integral calculus — *even differential equations!*

NEED TO EVALUATE functions, areas, volumes — solve equations — use curves, trig, polar coordinates — find limits for sequences and series? It's all here!

If you're in the biological, social or physical sciences, you'll be doing Bessel functions, carbon dating, Gompertz' growth curves, half-life, future value, marginal costs, motion, cooling, probability, pressure — and plenty more.

Important numerical algorithms are here, too: rational and Padé approximation, bracketing, continued fractions, Euler's method, Heun's method, iteration functions, Newton's method, predictor-corrector, successive substitutions, Simpson's method and synthetic division.

If you already have a scientific calculator, you can get just the *Calculus Guidebook.*

George McCarty

P.S. — YOU *CAN* DO IT! I'm certain I can show you this easy way to success with math, so ***I guarantee it to you unconditionally*** — if you are not satisfied just send it back for a prompt refund!

PHONE MC or VISA ORDERS TO
(714) 831-2637
Ask for our Catalog of books and calculators.

LOOK AT WHAT THEY SAY...

- *"...excellent diagrams and clear explanations."* —POPULAR SCIENCE
- Tektronix Engineer Bill Templeton says, *"The best, most clearly written book I have seen for improving your math skills."*
- *"...straight-to-the-point explanations..."* —HP Key Notes
- Professor H.I. Freedman of the U. of Alberta, states: *"There can be no question as to the usefulness of this book...lots of exercises...very clearly written...easy reading."*
- S.C. McCluney, Jr. of Philadelphia writes: *"With your book and a calculator the whole idea becomes clear in a moment, and is a MOST REFRESHING EXPERIENCE."*

TAX DEDUCTIBLE FOR PROFESSIONALS

YES, I want to learn math **fast** and **easy** with your *new method.* If I am not more than satisfied I can return it for a quick refund. RUSH me the guidebook or complete kit I select below (kit contains the guidebook and a modern TI-35 calculator).

☐ Send **FREE** Catalog of books & calculators

STATISTICS GUIDEBOOK		KIT
How many?	☐ $14.95	☐ $39.95
Handling/Shipping	1.00	2.00
Tax (within Calif.)	.90	2.40
TOTAL		

CALCULUS GUIDEBOOK		KIT
How many?	☐ $19.95	☐ $44.95
Handling/Shipping	1.00	2.00
Tax (within Calif.)	1.20	2.70
TOTAL		

For AIR SHIPMENT add $3.00 per item
Phone (714)831-2637 (MC or VISA only)

Enclosed is ☐ Check, ☐ M.O.; or charge my MC/VISA... exp._____

Card #_____

Name _____

Address _____

City_____

State, Zip _____

OUTSIDE USA, **surface shipping is $2.00 per book or $5.00 per kit. Air shipment is $5.00 per book or $10.00 per kit. U.S. FUNDS PLEASE.**

EduCALC Publications, Dept. A1
27953 Cabot Road
South Laguna, CA 92677

©1982

BOOKS

The real Albert Einstein, human clocks, amazing ammonites and megalith mania

by Philip Morrison

SUBTLE IS THE LORD . . . : THE SCIENCE AND THE LIFE OF ALBERT EINSTEIN, by Abraham Pais. Oxford University Press ($25). During Einstein's first visit to the U.S. in 1921 word reached him in Princeton that out at Mount Wilson some new measurements were showing a nonzero ether drift. His comment was: "Raffiniert ist der Herr Gott, aber boshaft ist er nicht." Subtle is the Lord, but malicious he is not. The remark is chiseled into stone at Princeton University still; Einstein interpreted his own saying when he consented to the inscription in 1930: "Nature hides its secret because of its essential loftiness, but not by means of ruse." Nor did the drift persist. Our current view of the vacuum, a complex virtual plenum with effects on particle energies and on cosmic expansion, "has nothing to do with the aether . . . , which is gone for good."

Einstein was an artist in exposition, as he himself said of Newton. His gift for vivid and apt expression was great; he was a "master of nuances" and a perceptive appraiser of the men and women of his time. "His portraits serve as the best foil for the opinion that he was a naive man." At ease only in German, he wrote every one of his scientific papers in that language. In this book we are given a sympathetic but clear-eyed view of his life and work and the physics he pursued from youth to death, examined mainly in his own plentiful papers, supplemented by a remarkably wide list of unusual sources.

Those include an unpublished "contribution to her brother's biography" written by his younger sister Maja in 1924, recent books from Germany and Switzerland, even Japan, an unpublished manuscript in the Morgan Library, reference letters from files in Zurich, Munich and Berlin, and of course the bulging Einstein Archives. The best of trails into the archives were opened by Einstein's secretary Helen Dukas. ("Dear Helen, thank you; it was wonderful." Miss Dukas died early in 1982, the most experienced of guides and witnesses to that life.) It is a rare treatment of a physicist dead for nearly three decades that can include entirely apropos references to current papers in *Physical Review Letters;* both the man studied and the man who studies him must be counted as unusual.

Professor Pais is a senior theorist, distinguished in his own right, who has in the past few years disclosed a flair for the history of ideas in physics. To this book he brings that power, and much more. He was lucky enough to walk and talk with Einstein at the Institute for Advanced Study in Princeton every few weeks for all the postwar years until a few months before the end. Vignettes of the old Einstein with the young quantum field theorist fresh from Bohr's institute in Copenhagen are poignant: once "he suddenly stopped, turned to me, and asked me if I really believed that the moon exists only if I look at it."

The last time, Pais was about to leave Princeton for a term's leave. Einstein, 75 years old, had been ill and away from his office for some weeks. It was natural to pay him a brief visit at home. "He was working. He put his pad aside at once. . . . We spent a pleasant half hour or so. . . . I told him I should not stay any longer. . . . I turned around as I opened the door. I saw him in his chair, his pad back on his lap, a pencil in his hand. . . . He was back at work."

It is the strength of this fine book that it displays and assesses that beloved work, in a thorough sampling of the *ix-erei* that is the terse language of the theorist. Always it appears in a prose frame of context and meaning; readers with some knowledge of this form of expression will be enriched by the brief treatments. Those who cannot grasp the x's and their subscripts are given much aid in learning the arguments and in sharing the life story that surrounds the equations. Among the many books on Einstein this one more than any centers on the work, its springs and its meanders, and of what came of it in the end in other hands—or has still to come.

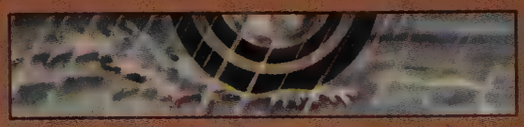
The events, public and private, of Einstein's long life are by now canonical, and yet both in facts and in understandings a good deal here gleams new and rings true. The baby was born with a strange skull shape (said grandmother, "Viel zu dick": Much too thick) and was slow in learning to speak. The grade school student brought home a fine report card; his mother wrote, her Albert seven and a half, "He was again number one, his report card was brilliant." Throughout the years in Luitpold Gymnasium the boy earned either the highest or the next-highest mark both in mathematics and in Latin. The poor-student myth is unfounded, although the independent Einstein himself recounted that a high school teacher objected to his presence. "You sit there in the back row and smile, and that violates the feeling of respect which a teacher needs from his class." Einstein was fascinated by experiments, and from his first Zurich years he planned big ones, for which his professor was not enthusiastic.

Einstein's experiments and patents, sought with able partners, often on their initiative, make by no means a petty list, from the first successful demonstration that magnetized currents implied rotation to the electromagnetic pump for home refrigerators, now found in nuclear reactors. Twice married, first to a fellow student at Zurich, later to a devoted and motherly cousin, he was no model husband and father, even in his own eyes. When his friend Besso died, old Einstein wrote: "What I most admired in him as a human being is the fact that he managed to live for many years not only in peace but also in lasting harmony with a woman—an undertaking in which I twice failed rather disgracefully."

It is fascinating to follow the progression of his researches. Pais remarks that his early career reminds one of Mozart: effortless, fresh, a flood of successes. Then the chronicle rather passes into the mode of a Beethoven, persisting, learning, maturing at last to peerless mastery: general relativity and its cosmology stated with clarity by 1918. If one inference is to be drawn from Pais's incisive study, it is that the lonely excursion Einstein took, a man apart, into the unified field theory after 1925 to his last years was not chiefly a search for a single theory of forces. That is a current quest of physics. His was a grand effort to conjure the quantum graininess out of the flowing field world, to make particles and probability necessary somehow by the demands of invariance on the nonlinear but strictly causal fields gravity required. We do not yet know how that drama will end; the unification of today, brilliant as it is, remains incomplete. Theorists do not at all share Einstein's deepest goal of restoring Newtonian causality to a world reliably ruled to a good many decimal places by the subtle prescriptions of quantum fields.

It is tempting to imagine for each of

FOR THE WET



FOR THE WHITE

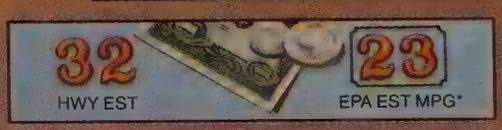
FOR THE WINDING

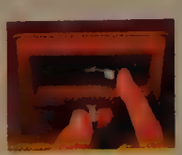
FOR THE WILD

FOR THE WARM

FOR THE WEEKEND

FOR THE WALLET





W is for Eagle Wagon. And for the wherever. Because Eagle is the only automobile in the world that lets you go from 2-wheel drive to full-time 4-wheel drive at the flick of a switch. 2-wheel drive for high mileage on the highway. 4-wheel drive for worry-free traction and security when the roads weave, the weather worsens, or when you just want to wander where there is no road.

And Eagle's well-appointed interior makes the going all the more wonderful. Why not see the 1983 Eagle Wagon today. Wherever you're going, whatever the weather, it'll fit you to a **W.**

The 2-wheel/4-wheel drive EAGLE

FROM AMERICAN MOTORS

*Optional 5-speed stick. Use these figures for comparison. Your results may differ due to driving speed, weather conditions and trip length. Highway and California figures lower.

THE FIRST PERSONAL COMPUTER

When you sit down with the Professional from Digital Equipment Corporation, you discover something very quickly: this is not just another personal computer.

It is a thoroughly professional system designed for the sole purpose of working in a professional, business environment. Which is why it has capabilities superior to other personal computers you might have seen before.

EASIER TO WORK WITH.

Through Computer-Based Instruction, you start using the Professional™ computer the same day you get it. Further, with software that's certified by Digital, the Professional computer gives you one consistent menu for all your applications. A first in personal computers. There's also a HELP command to answer questions about operation. And if something goes wrong with the system, it will actually draw you a picture of the problem and tell you what to do about it.

BETTER COMMUNICATIONS.

The pictures and graphics you see on the Professional are four times sharper than conventional personal computers. The text can be displayed in bold, underlined, double-sized, or practically any other form that will make communications clear. And the unique Telephone Management System opens the way to

Dow Jones News/Retrieval is a registered trademark of Dow Jones & Co., Inc.
Professional is a registered trademark of Digital Equipment Corporation.

AS PROFESSIONAL AS YOU ARE.

voice communications in personal computers for the first time.

MORE INFORMATION.

The Professional computer can hold a staggering 5.8 million bytes of information right at your desk. When you need information from other sources, you can be networked with other Professionals, larger systems from Digital, IBM mainframes, and data services like Dow Jones News/Retrieval.*

And the Professional computer can do *more* with that information. Its "multitasking" capability means it can work on one or more jobs while you work on yet another.

MORE COMPUTER.

Where other personal computers have reached the limit of what they can do, the Professional computer is only beginning. It has true minicomputer power, and thus offers extraordinary growth potential. It has many capabilities that are only starting to be explored.

We'd like you to meet the Professional computer. In person.

To find out where you can see Digital's Professional™ 300 Series computer, call 1-800-DIGITAL. Or write: Digital Equipment Corporation, 129 Parker Street, Media Response Manager PK32/M94, Maynard, MA 01754.

digital

MORE PERSONAL. MORE COMPUTER.

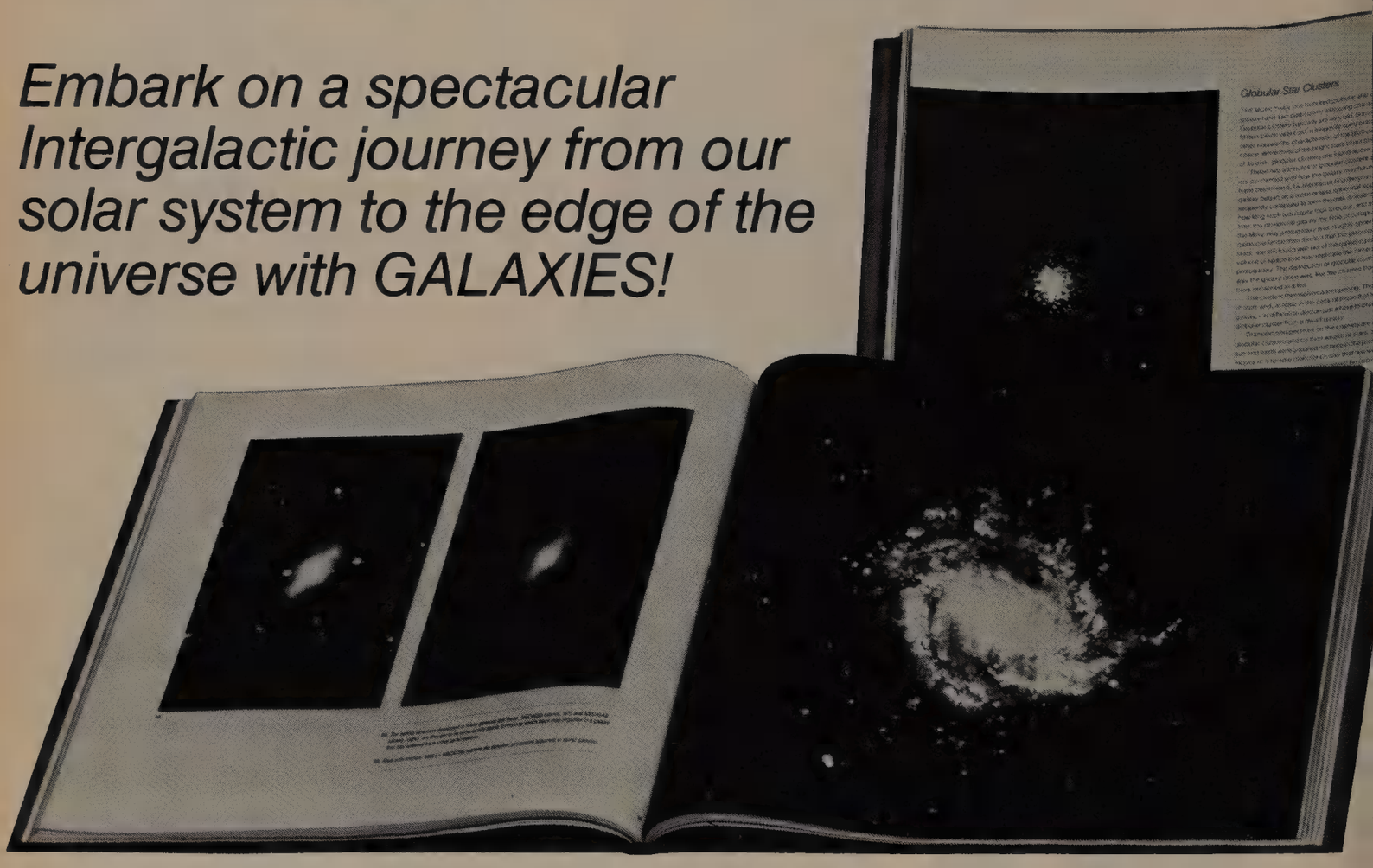
Now, you are invited to take this magnificent reference an

GALAXIES for only

as your introductory selection to the ASTRONOM

You simply agree to buy 3 more books—at handsome discounts—over the next 12 months

Embark on a spectacular Intergalactic journey from our solar system to the edge of the universe with GALAXIES!



Shooting stars...novas...star clusters...black holes and much more await you in the inspiring and beautiful pages of GALAXIES.

With this exciting offer, you can take GALAXIES, a $75.00 value, for only $2.95 as your remarkable introduction to the Astronomy Book Club. Handsomely bound in the toughest, most durable materials and printed on high quality paper, your impressive volume has been constructed to last a lifetime.

And GALAXIES is visually stunning. It includes a virtual treasury of outstanding photography—each a work of art in itself—all collected by astronomers from observatories throughout the world.

In GALAXIES, 145 spectacular photographs (39 in blazing full color) take you on an intergalactic journey beyond the realm of the ordinary...a fantastic journey to the outer reaches of the universe.

GALAXIES is more, too. It includes 20 magnificent line drawings by noted illustrator Sarah Landry and a huge 13" x 56" gatefold photographic mosaic of the Milky Way.

GALAXIES is the most beautiful and significant book ever produced on the study of the worlds beyond our own...a book you will be justifiably proud to own.

In fact, Sky and Telescope magazine had this to say about GALAXIES:

"The artwork that accompanies the text and photographs is superb. The paper and binding are of high quality... a marvelous gift for someone who wants to know about the universe."

adventure volume...

$2.95

BOOK CLUB.

GALAXIES

the most evocatively beautiful volume on the nature of deep-sky splendors ever published.

- An authoritative reference work...and an exciting literary adventure
- Spans an extraordinary 13" x 14½"...182 pages of vivid text and clear, concise captions
- Tough, durable binding and lustrous, coated stock, smudge-proof paper
- 106 dramatic duotone photos...39 full color photos—many full page size
- Original hardcover edition
- Written by award-winning author, Timothy Ferris
- 20 line drawings by noted illustrator, Sarah Landry
- Includes extensive bibliography, glossary and index
- A $75.00 value—yours for only $2.95 as an introduction to the Astronomy Book Club

"...a very good candidate for the most beautiful book in the world... I was absolutely overwhelmed by its excellence—text *as well as* photographs."

—Isaac Asimov

According to the New York Times, "The book is at once a linear, visual, historical, futurist, cultural and philosophical journey to forever."

And now you may have the original hardcover edition of GALAXIES for the exceptional bargain price of only $2.95 (you save over $72.00 off the publisher's price) as your introduction to the additional delights that await you as a member of the Astronomy Book Club.

Astronomy Book Club is a very special club—one that is operated exclusively for astronomy enthusiasts. Recent offerings include titles by such eminent authorities as Carl Sagan, Robert Jastrow, John Gribbin and Paolo Maffei. The Club is devoted to making the finest astronomy books available conveniently and inexpensively!

4 Good Reasons to Join

1. The Finest Books. Of the hundreds and hundreds of books submitted to us each year, only the very finest are selected and offered. Moreover, our books are always of equal quality to publishers' editions, **never** economy editions.

2. Big Savings. In addition to getting GALAXIES for only $2.95 when you join, you keep saving substantially—up to 30% and occasionally even more. (For example, your total savings as a trial member—including this introductory offer—can easily be over 50%. That's like getting every other book free!)

3. Bonus Books. Also, you will immediately become eligible to participate in our Bonus Book Plan, with savings up to 70% off the publishers' prices.

4. Convenient Service. At 3–4 week intervals (16 times per year) you will receive the Book Club News, describing the Main Selection and Alternate Selections, together with a dated reply card. If you want the Main Selection, do nothing and it will be sent to you automatically. If you prefer another selection, or no book at all, simply indicate your choice on the card, and return it by the date specified. You will have at least 10 days to decide. If, because of late mail delivery of the News, you should receive a book you do not want, we guarantee return postage.

If reply card has been removed, please write to:
Astronomy Book Club, Dept. L-AW4, Riverside, N.J. 08075
to obtain membership information and an application.

SCIENCE/SCOPE

High efficiency solar cells now being developed could cut substantially the weight and area of some satellite solar panels. These cells, made with gallium arsenide by a special liquid-phase epitaxial growth process developed by Hughes Aircraft Company scientists, have shown efficiencies higher than 19% in converting sunlight into electricity. This compares with efficiencies between 15% and 16% for the best available silicon solar cells. In addition, the gallium arsenide cells can operate at higher temperatures and can tolerate more high-energy proton irradiation in space than silicon solar cells.

A new wireless entertainment system will use infrared light to carry music and movie soundtracks to passengers on commercial and corporate jet aircraft. The system, under development at Hughes, transmits a digital infrared signal that is received and decoded by a passenger's headset. The signal is completely harmless to people and won't interfere with other aircraft equipment. The infrared system transmits 16 channels. It would weigh less than half of a conventional wired system and would cost a third less. Since there are no wires with this system, it is particularly suited for aircraft with changeable seating.

Communications among Brazil's cities and remote villages will be improved when the first domestic communications satellite system in Latin America goes into operation in 1985. The primary function of the two Brazilian satellites will be to improve telephone and television services. The spacecraft are based on a Hughes model that is the world's most widely purchased commercial communications satellite. Now 27 are under contract. Hughes will serve as major subcontractor to Spar Aerospace of Canada. Embratel, Brazil's government-owned telecommunications company, will buy and operate the satellite system.

Now in its 17th year, Science/Scope remains among the best-read corporate advertising campaigns in history. The campaign was created to explain what Hughes does and how. About 160 different ads are produced each year, with a mix of over 175 paragraphs for use in over 80 publications serving over 60 nations. It is translated into Arabic, Chinese, French, German, Italian, Japanese, Korean, Portuguese, and Spanish -- 10 languages in all, including English. It consistently scores in the top 5% to 10% in readership surveys.

Career growth opportunities exist at all levels at Hughes Support Systems for a variety of engineers qualified by degree or extensive work experience. They include systems engineers, applications engineers, software and hardware design engineers for major simulation and test equipment programs, and automatic test equipment engineers. Also, field engineering posts throughout the U.S. and the world offer travel, autonomy, and responsibility. Call collect (213) 513-5238. Or send your resume to Professional Employment, Dept. SE, Hughes Aircraft Company, P.O. Box 9399, Long Beach, CA 90810-0463. Equal opportunity employer.

Creating a new world with electronics

HUGHES

HUGHES AIRCRAFT COMPANY
CULVER CITY, CALIFORNIA 90230

(213) 670-1515 EXTENSION 5964

Einstein's individual victories the effect that removing that single most productive theorist of this century might have had on the history of physics. His first fine work of all, on the reality of molecules, was foreshadowed, in part repeated independently, by Willard Gibbs and Marian Smoluchowski. The theory of special relativity lay a little beyond the reach of Henri Poincaré, who had recognized both the technical transformations needed and the requirement for a rethinking of simultaneity but could not fuse them. Would another young physicist have come along to do that? It is a good bet either way.

The theory of gravitation required still more insight. There were signs, however, that among the mathematicians of Göttingen around David Hilbert that theory too might have emerged. (Given, to be sure, the hints of Einstein's earlier approaches, Hilbert did write the equations correctly quite on his own.) Perhaps it was the photon, the particlelike bearer of Maxwellian waves, that most needed Einstein himself. From his 1905 photoelectric effect straight through a series of permanently valuable and brilliant inferences about radiation and matter, Einstein remained loyal to the reality of the photon, which he perhaps more than anyone established. Even Bohr once abandoned the photon, when in 1924 he despaired of individual atomic acts of conservation of energy and momentum.

Einstein remained steadfast; much earlier, in 1909, he had written: "It is my opinion that the next phase ... will bring us a theory of light that can be interpreted as a kind of fusion of the wave and the emission theory ... [which] are not to be considered as mutually incompatible." When that theory appeared in full in about 1930, Einstein found it repugnant. He maintained it was a deceit. No final truth was in it, for all its predictive success. Pais senses a reason for this stubborn philosophical set on the part of the very man who argued the world into accepting not only a new structure for space and time but also a new entity at once a wave and a particle. Einstein almost said it himself: "To the discoverer the creations of his imagination appear so necessary and so natural that he is apt to treat them not as the creations of his thoughts but as given reality." Twice he had found reality in depth by insisting on the constraints imposed by certain freely chosen simplicities. A third time has not come.

"Einstein was the freest man I have known." He was "the master of his own destiny," free to ask, fearless of time and death. No tragedy lies in his attitude toward the quantum theory or in his unsuccessful search for the unified field. So writes Pais, physicist and friend. The tragedy of Einstein is not personal but universally shared; we are linked through the death camps, the shadow of the distant mushroom cloud, the endemic racist plagues. Einstein free is symbol of this ironic century of ours, its sands not yet run out.

THE CLOCKS THAT TIME US: PHYSIOLOGY OF THE CIRCADIAN TIMING SYSTEM, by Martin C. Moore-Ede, Frank M. Sulzman and Charles A. Fuller. Harvard University Press ($25). The storyteller vividly reminded us that to "the cat that walks by itself" all places are alike. Here a striking factual diagram documents even more: for that cat all *times* are alike as well. A month of plotted activity hour by hour shows little temporal pattern. Bursts of activity and rest follow each other irregularly for a couple of hours, giving little sign of any self-sustained longer rhythm, independent of even a shifting cycle of light and dark. Mice and men are not like cats. Both species show strong activity rhythms. Living without time cues in a carefully isolated laboratory, house mice and human beings both show a regular free-running circadian pattern, close to but not matching the 24-hour rhythm of earth spin. (Cats and dogs and bobcats are exceptional.) Most mammals—and living forms in general—have a marked built-in circadian rhythm, with a period that differs among individuals and from species to species. House mice tend to run fast: their stable free-running period centers near 23.5 hours. Human beings run a little slow: their internal-activity clocks average about 25 hours around the dial.

This carefully documented book is a comprehensive yet nontechnical introduction to a fascinating new system in human physiology: the internal timing system, tightly bound to sleeping and waking, plainly part of the still unexplored richness of the unconscious brain-body connection. The volume develops the ideas clearly and logically, with a minimum of mathematics and many graphs from the big body of experimental results. Its clear aim is to draw everyday and medical implications from the human case of this subtle outcome of life's rotating environment.

We live by the clock, either by man-made dials or by the older one of dawn and dusk. Many properties of the body, not only our most conspicuous rhythm of sleep and wakefulness, follow the day closely. Both internal processes and the response of the system to all kinds of outside effects show strong diurnal patterns. In the laboratory mouse, for example, clear cyclical effects of the time of day have been demonstrated among 50 or 60 distinct physiological variables.

One plot shows a sample of human daily patterns, varying sharply in phase and form among the quantities measured. Sleep comes by night, and with it comes the maximum blood content of growth hormone but the minimum of internal temperature and urinary potassium excretion. Cortisol levels are at a minimum in the blood as sleep encroaches but wax to a maximum shortly before waking. Evidently the conscientious intern who watches all night over her patient's measured states must take account of such natural waves of activity. These patterns mature; infants sleep and wake as haphazardly as cats for the first month or so (on a demand feeding schedule). Then a marked rhythm sets in that runs a little slow of the dawn-dusk cycle, until after four months of life the external world of day-night and of human activity in general has entrained and synchronized the biddable inner clock.

It is the controlled study of that entrainment by external change that best reveals novelty amidst all these harmonies. The outside pacemaker can cue time commands at will up to a point, usually by a few hours, plus or minus, away from the inbuilt period. The cue need only be small; a single brief light pulse once every 24 hours is sufficient to synchronize the squirrel monkey, our little diurnal primate cousin. The spontaneous loss of that synchronization, once all external pacemaking is removed, is the most remarkable recent result, particularly in human beings.

One typical subject showed a steady 24-hour day-night cycle, still entrained to the stars, for the first five days of total laboratory isolation in time. Then an inner clock took over that kept a 25-hour day. That clock was steady, and both the rest-activity cycle and the body-temperature cycle followed it closely. At about five weeks something inside changed. Thereafter the body-temperature cycle shortened a little, although it was still a bit slow on the 24-hour day. The sleep cycle, however, lengthened markedly. It kept its own time, with a regular if intricate structure of long and short sleeps, on an inner day that was now more than 30 hours long.

We must have two chief timekeepers, not one. Their mutual interaction is well described by a simple mathematical model. One clock controls the skin temperature and many related rhythms, the other controls the sleep-wake cycle and the many other functions that derive from sleep. There is even strong evidence, not yet certain in human beings, that serves to place the two clocks. One of them, an organ we share with most other mammals, is seated in a pair of millimeter-size clusters of special neurons in the hypothalamus deep within the brain. A long history of surgical

excision in monkeys has made this site pretty sure. Clinicians have for decades realized that brain tumors near the corresponding structures in human beings can bring profound sleep-wake disorders. More remarkable still, in monkeys (not, of course, in human patients) quick autopsy after the injection of a radioactive tracer into the retina has disclosed an active tract of neural fibers leading directly from the retina to that clock center. We are here close to an understanding of how we wake by dawn.

The other timekeeper is less well known; there is evidence that it too lies in the hypothalamus, near but distinct from the other one. The retinal-linked clock times the skin temperature, the growth hormone of the blood and our dreamless form of sleep. It runs freely at a very slow rate, although it normally senses night and day. The 25-hour clock controls the rapid-eye-movement sleep of dream, potassium excretion in the urine and the generation of heat deep within the body. Its rhythm carries over to the other slower oscillator, which in return feeds back strong signals of its own informed vision of light and dark. Together they fix our usual coordinated 24-hour rhythm. Of course, this account is much elaborated in the full text by these three experienced research workers. They—a physiologist, a biologist and a neuroscientist—have worked together fruitfully for seven years on these inward beats of night and day.

We live now, as we have ever since the discovery of fire, increasingly cut off from earth spin. Daylight-saving time, shift work and swift jet travel across time zones are clear examples of acute stress on the ancient timing system. Consider too the inhabitants of the Arctic who have chronically endured winter months without the unfailing dawn cues of the ancestral tropical latitudes of our hominid kind. There are gloomy tales: once an airline pilot fell asleep in the very act of touchdown, 200 feet above the runway. (The copilot was awake.) In 1978 a jet overflew Los Angeles on autopilot by 100 miles; the entire flight crew was asleep. The Three Mile Island power reactor was operated by a crew that rotated shifts around the clock on a weekly basis, conventional enough but much too fast for good adjustment. The Navy judges that eight-hour shifts are too arduous for the crew of a missile submarine. Since the boat can hold only three shifts of men, however, it operates on a stressful 18-hour day, quite unphysiologically.

The trick for time-estranged travelers is probably to avoid shifting phase at all when the visit is brief enough. If a shift is necessary, experience and the beginnings of understanding favor a delay of phase over an advance. It is easier to entrain our slow double inner clock in days that are longer by a couple of hours than in days that are shorter; a westbound time gain is more quickly compensated than an eastbound loss; certain out-of-phase insomnias have responded to a careful regimen of 27-hour days. "The study of the circadian timing system is thus at a fascinating stage"; here is human physiology entering a completely new dimension.

THE AMMONITES: THEIR LIFE AND THEIR WORLD, by Ulrich Lehmann. Cambridge University Press ($19.95). THE AUDUBON SOCIETY FIELD GUIDE TO NORTH AMERICAN FOSSILS, by Ida Thompson. Alfred A. Knopf, Inc. ($12.50). Three or four hundred meters deep in the dark sea from Fiji to the Philippines the pearly chambered nautilus hunts by night. Slowly swimming backward, the creature's propelling jet spurts from its head past 90-odd little tentacle arms. The nautilus is an antique like the coelacanth, not much changed over an ancestry going back proudly to the Ordovician. Before ever fish swam these animals and their earlier nautilid kin ruled the seas, their chambered shells often not yet coiled.

The upland limestones of southwestern Germany yield up many fossil coils that resemble the beautiful planar spiral chambers of the nautilus. So widespread is one early form that it has been proposed as the German heraldic figure. Sometimes the calcareous shell material is itself preserved, with its pearly sheen and elegant form plainly "among the most beautiful fossils in existence." Commoner are clear interior molds, all complex detail visible.

These remains have been collected and studied for a long time in the classic localities. There is such a fossil engraved by a Paleolithic artist, from a cave near Ulm. Many thousands of forms are now recorded worldwide, from submillimeter size to nearly two meters across, spanning a time of more than 100 million years, until they end along with the dinosaurs at the Cretaceous-Tertiary time boundary. For a century the intricate morphology of these abundant remains supplied in all their variety the basis of a stratigraphic dating procedure, but the biology of the animals themselves was neglected in the "somewhat schematic concept of index fossils." Indeed, the living nautilus is itself little known, only recently observed in captivity.

Professor Lehmann's book, a broad and well-illustrated introduction to the life that secreted the spirals, is the work of a paleobiologist who has made a major contribution to the newer understanding. The intricacy of the shapes that must be understood in three dimensions and the general strangeness of molluscan life in the sea do not allow easy grasp by the general reader, even though no previous special study is assumed. The sense of excitement in the new inferences drawn from recent finds nonetheless comes across sharply; the brief book will repay the study it needs to anyone who enjoys the paleontologist's hunt for meaning.

The key to the life of the ammonite animal was the finding of its jaws. For a long time it was thought ammonites had none, no analogue to the strong, sharp parrot beak of the nautilus, made of resistant chitin-reinforced calcite. Modern cephalopod jaws are abundant on today's ocean floor, here and there even densely covering the bottom. Only in 1965, however, were the jaws of an ammonite clearly recognized, in a well-preserved specimen collected by students at the University of Montevideo. The find was unique in its preservation; the phosphate nodule had only to be cracked open to reveal to the knowing paleontologist (Professor Closs of Puerto Alegre in Brazil had his Ph.D. from Tübingen for work on certain curious double calcite structures found isolated or near ammonites) the entire relevant apparatus of a mollusk, a rasplike radula between two slightly flattened jaws.

Within 10 years the riddle was pretty well answered, although there remains a question or two. Professor Lehmann, the author of *The Ammonites,* ground away the calcite tenth-millimeter by tenth-millimeter to follow and map the darker contours within the white material that filled the body chamber of one suitable ammonite specimen. His first hint was a structure closing the aperture of the body chamber. Within the familiar but long uninterpretable structure there was at last disclosed a second pair of fitted lobes; it was unmistakably a pair of jaws made of a "horny-coaly substance." A three-dimensional model was patiently prepared.

Teeth and jaws offered new meaning. The ammonite's blunt-edged lower jaw was not a cutting tool or even a sharp beak. The jaw muscles were not strong. The creature held its jaw like a shovel just above the sea bottom as it floated slowly along, feeding on bottom organisms. There is much corroboration by now; Lehmann found "a small insignificant-looking black fragment" in a London collection that yielded a few half-inch spirals. In them pairs of jaws were to be found; in one even a stomach holding a last meal of small foraminifera and ostracods.

Those ornate spirals and the many aberrant forms that look more like saxophones than like French horns are now newly plausible; shell form was not significant ecologically. Water resistance

Home sweet Honda.

The 1983 Accord Hatchback offers all the comforts of home. Plus a few more of its own.

The Accord's spacious interior and plush trim are wonderfully inviting. And once on the road, its excellent handling and roadability are proven at every turn.

Quality Honda engineering features include front-wheel drive, 4-wheel independent suspension and power-assisted self-adjusting ventilated front disc brakes. Steel-belted radial tires, a maintenance reminder and electronic warning system all come standard.

A new 4-speed automatic transmission with variable-assist power steering is also available. It not only makes driving simpler, but offers improved automatic fuel economy as well*.

The Accord Hatchback can improve your personal economy, too. Because in spite of all it has to offer, this hatchback is still our lowest priced Accord. Which is one more reason so many people find it so nice to come home to.

*For 5-speed transmission, 32 EPA estimated mpg, 45 estimated highway. For automatic transmission, 29 EPA estimated mpg, 40 estimated highway. Use estimated mpg for comparison. Your mileage may vary according to weather, speed or length of trip. And you can expect actual highway mileage to be less. California mileage will be lower.

©1982 American Honda Motor Co., Inc.

HONDA

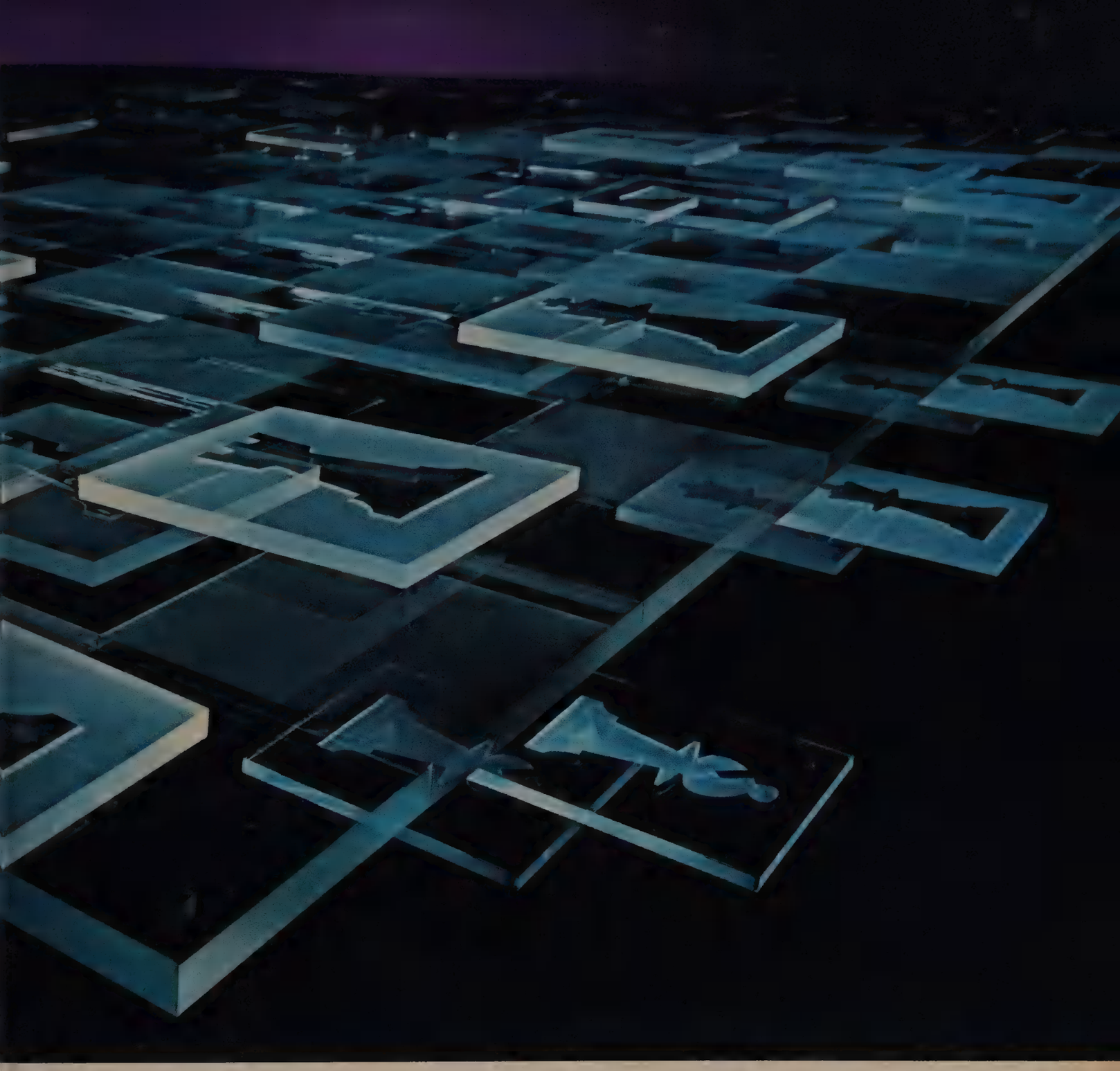
We make it simple.

An IBM 4341 super-mini: chip design at Hughes

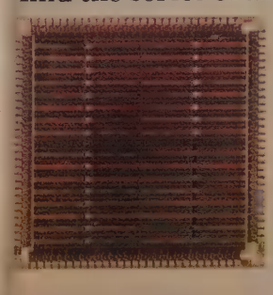
The board's half the size of a thumbnail, with 100,000 pieces. And it's your move.

Finding paths for new interconnections on an integrated circuit is like no chess game you have ever seen. The first few moves are easy, but with up to 100,000 electronic devices crammed onto a minuscule silicon chip, the possibilities soon become astronomical. That's why engineers at Hughes Solid State Products in Newport Beach, California use an IBM 4341 super-mini.

Brian Tien, head of design automation, says, "Without the IBM 4341 we couldn't get this much function on a chip. With it, we can finish a circuit in a few weeks instead of many months."



The engineers work interactively with the system, assigning logical functions to devices on the chip. Then, using software, they route conductors—deposited strips of metal—to connect the logic gates. If a pathway becomes too crowded, another layout is automatically provided. And another. Until they find the series of moves that works.



This Hughes integrated circuit is a high-density gate array with two-level metal interconnections. Actual size is only 0.3 inches square.

They can think about the problem, and not worry about the computer, with microcode-assisted IBM software called VM/CMS. "Our people find VM ideal for interactive computing," Tien says. "As many as 40 design engineers make demanding use of the 4341 at once. Response time is excellent. The full-screen editor speeds up programming. And the executive language is simple, yet powerful."

The 4341 is an excellent example of IBM technological leadership. Internal data paths, and the arithmetic and logic unit, are 64-bits wide, built of 64K-bit chips that IBM has been mass producing since 1978. Real memory goes to 16 megabytes. The multiple, semi-autonomous processing units use high-density, large-scale-integration technology.

IBM offers engineering and scientific users extensive support: consultants, education and access to professionals. Tap into our many years of experience.

For an informative brochure, write Dr. Jack W. Hugus, IBM Engineering and Scientific Marketing, 1133 Westchester Avenue, White Plains, NY 10604.

IBM®

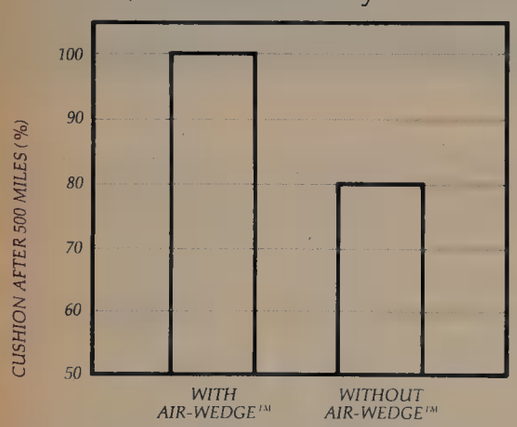
THIS NEW AIR-WEDGE IS OVER 10,000 MILES LONG.

It's hard to believe. But we just made something shorter without reducing its length.

Remember the NIKE-Air™ sole? Well, we just came out with a condensed version. The Air-Wedge.™

We did it the minute we learned from our survey that three out of four runners were heel strikers. With the Air-Wedge, we can now give those runners the protection of Air, but only where they need it most. Right under the heel.

True, the Air-Wedge is a lot shorter. But it reaches just as far—about 10,000 miles actually.



After 500 miles, shoes with the Air-Wedge lose none of their cushioning properties. Those with EVA foams, however, will suffer a 15% to 20% loss.

And probably even further. We don't know how much further because our lab technicians threw in the towel after watching the NIKE-Air sole successfully handle over 6,000,000 impacts.

That's the funny thing about Air. You'd think it would be fragile, susceptible to blow-outs, leaks, etc. When in reality it will outlast virtually every other part of the shoe, from the laces to the outsole.

That's not even the good news.

What's truly phenomenal about the Air-Wedge is that its cushion won't break down. It will absorb just as much shock on the first step as it will on the 5,999,999th.

Pretty remarkable. Especially considering the typical shoe with an EVA wedge will lose about 15 to 20 percent of its cushion after just 500 miles.

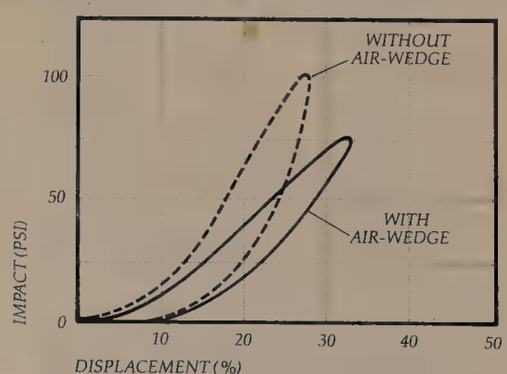
Such rapid compaction isn't just sad, it could be dangerous.

Because, if you're a runner, you might as well hit your heel with a five pound hammer. That's how much shock is generated with every step.

Not so funny.

And, unless you like buying a new pair of shoes every 300 miles, not so cheap.

But the Air-Wedge does more than just keep its cushion. According to tests carried out in our sport research lab in Exeter, New Hampshire, it gives better cushion the first day out of the box.



When subjected to impact testing, a new shoe with the Air-Wedge doesn't bottom out like a new shoe with EVA foam. Instead, it responds in a more linear fashion, delivering 12% better cushion. The steeper the slope on the curve, the harder the material becomes when it is compressed under your heel.

Whenever we stuck an Air-Wedge in a developmental shoe, shock absorption immediately jumped 12 percent. The reason being, the Air-Wedge simply doesn't bottom out, or get stiffer as it is compressed.

Okay, but how stable is it? Answer: surprisingly good. We also ran some direct comparison tests between two similar shoes, one with the Air-Wedge, the other with an EVA foam wedge. As far as controlling rearfoot motion, they both did the job equally well.

One last thing. Not all runners will run on the same amount of Air. We have scaled the pressure in the Air-Wedge according to shoe size, so that runners of all sizes will receive the appropriate amount of cushion.

Pegasus

Now where can you find this amazing device? In two new Nike models. The Pegasus. And the Odyssey.

We've made many claims about our NIKE-Air shoes. And with the introduction of the Air-Wedge, we're making even more.

But, then, we know it's going to let you down.

A lot easier. And a lot longer.

NIKE

Beaverton, Oregon

meant little to the slowly drifting animals, mere living dustpans over the bottom. Ammonites dwelt on the continental shelf, acting as health police of the sea floor. They seem to have filled the role of marine snails today, consumers of all small morsels, dead or alive. There is a South Dakota ammonite with 16 holes in its shell in a pattern neatly fitting the teeth of a marine reptile of the period. Other coils are cut away as though with scissors by some large crab. But they came to an end, those neutrally buoyant, underpowered subs, once the powerful bony fishes and newly instrumented unshelled cephalopods arose, probably less from direct predation on the armored adult coils than from the loss of tiny larval forms and the tough competition overall.

The jacket strip placed around the second of these books, a chunky little field guide, the right size to stuff determinedly into an outside pocket, is adorned by a pearly four-inch ammonite. That gleaming coil was found long ago in South Dakota and is held by the Peabody Museum at Yale, where lately it served as model for one of the striking photographs that are the heart of the work. From the Japanese printers through the designer and the photographer to the author, a Rutgers paleontologist, the entire team merits readers and praise.

This is a first-rate entry to the shelf of field guides for the American amateur naturalist, young and old. Its apparatus is as good as its images: the way to identify a find or cherish a hope is to seek its general form in the silhouettes, turn to the indicated color plates to find similar examples, narrow in by using the geologic period of your location judged from the detailed maps given here region by region and choose the best match. The text paragraphs help you to decide and enrich your understanding of the ancient plant or animal. Glacial, granite New England and Precambrian Quebec are held not sufficiently fossiliferous to appear on these maps, but people there will find the book an incentive to travel, if perhaps only to a museum. The specimens photographed come from one of four distinguished Eastern museum collections, and they are well identified and located. The book shines with care and forethought.

MEGALITHOMANIA: ARTISTS, ANTIQUARIANS AND ARCHAEOLOGISTS AT THE OLD STONE MONUMENTS, by John Michell. Cornell University Press ($16.95). Fierce Farmer Robinson, the churlish "stone-killer," wantonly broke up Avebury stones by pouring water over each rock as it lay across a pit filled with a blazing wood fire. Robinson's face and name survive because he was made the subject of the tailpiece of the Avebury book of the Lincolnshire physician William Stukeley. Topographic engraver of genius, devotee of the English countryside, careful recorder of its monuments, Stukeley was caught up in the 1740's by an inner religious conversion, out of which flowed his two volumes on Stonehenge and Avebury, celebrating his conviction that it was the British Druid priests whose noble serpentine temples are now the standing stones and artificial hills of Wiltshire. The poet-prophet William Blake followed but inverted Stukeley's generous vision: "All things Begin and End in Albion's Ancient Druid Rocky Shore." You can see Blake's own trilithon engraved amid moon and stars against these words, from his *Milton;* cruel priests at Stonehenge had for Blake usurped the ancient humane cults.

John Michell has here assembled a marvelous sampling, mostly in period drawings and old photographs, to give an account of the vicissitudes the megaliths of Britain and Atlantic France have undergone. That damage came physically less from lithoclasts such as Robinson than from zealous excavators, and in symbol less from evocative poet and artist than from those who over the centuries have sought to draw meaning from the structures in a more argumentative style. Once General Pitt-Rivers excavated a barrow in style nothing was left of it. There is little doubt that the projection of attitude and theory on mute stones, the elevation of modestly supported conjecture to high dogmatic truth, remains a vivid part of the encounter of each generation with the enigma of purpose that is still dim behind circles and dolmens as plain as a pikestaff.

Stukeley's images were pretty faithful to the views he saw. The iconography of Stonehenge is rich, from a 14th-century manuscript rendering of the stones as a *rectangular* temple to the painterly visions of Constable and Turner. Michell clearly sympathizes with the visionaries more than with the meticulous archaeological excavators; their century of the spade ended the Druidic mythmaking and the Roman temples, but it carried with it a certain disdain for the great circles themselves, a disdain that extended to the denigration of all efforts to seek their meaning.

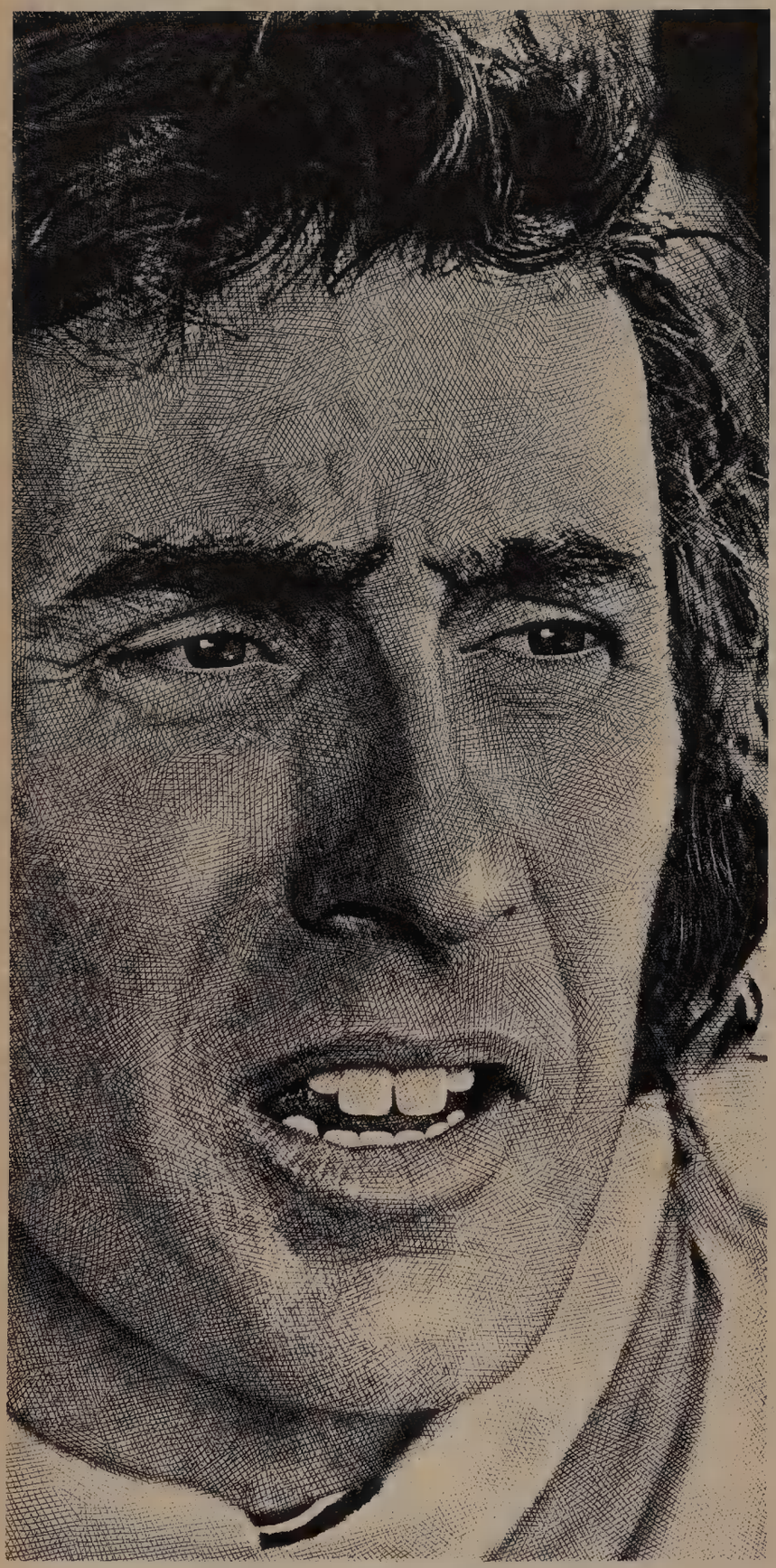
Stonehenge as a whole received little scholarly attention, even though from old Stukeley to the careful measurements by the astronomer Norman Lockyer there were plenty of hard data to link the circles in some part with orientation to sun and stars. It is less than 20 years since the arguments of Gerald Hawkins and the measurements of Alexander Thom—by now meeting their newly skeptical reaction—ignited anew the fires of learned imagination. They began a lasting rehabilitation of the idea that astronomical elements at least in part inform the circles and the alignments.

Holes and spirals and lines on the rocks are nowadays read as diagrams, even as inscriptions. There are a delightful few pages on the pitfalls of decipherment, reaching sometimes "the very heights of inspired lunacy." In 1833 an Icelandic scholar construed five heroic poems in an ancient meter out of the marks on a long-celebrated Swedish rock. It has turned out, alas, that the marks so cunningly read are natural lines and cracks, although his poem sequence remains a masterpiece of the old form. "No one has ever disputed its fine quality."

Michell is more willing than most scientific readers can ever be to consider the suggestion that the sites of these antique wonders are in fact special geophysical locales, magnetic or healthful in some way divined by ancient and modern dowsers and geomancers, with or without old visions or splendid new instruments. Surely we see here, as Michell himself wisely points out, the rise of a new myth fit for our ecologically concerned times, the search for lore that once fitted our yearning species more harmoniously into our lands. Pierced stones natural and manmade from Britain to India hold folk promise for health and fertility; infants are still passed through them.

The symbol of birth is clear enough without evoking some unlikely energy fields; one doubts the need for epidemiological studies of the neighborhood. The book nonetheless does much in a gentle way to document the perils of grand conclusions from slight premises, and the enduring danger of the very same preconceptions that are indispensable to progress. Let a hundred schools contend, says the author; he mostly wants to infect the reader who has not seen those "battered old rocks and mounds of earth" with the appealing half obsession he calls megalithomania. Magical cairns are not unknown on the New World plains, either; a wonderful drawing of a buffalo skull placed on a pointed rock as it was found by Prince Maximilian zu Wied on his celebrated expedition to the Yellowstone in 1841 and another of the winding Great Serpent Mound of Ohio help to Americanize this delightful book, if with only a few images. The epigraph is a pleasure; the hard-boiled editor of *Antiquity* remarked in that journal in 1961: "The problem in archaeology is when to stop laughing." This reader still laughs at any claims of efficacious magnetic structures, but the notion of antique sight lines across country is slowly becoming less risible.

Who needs criticism?



Jackie Stewart Three-time world champion driver, now consultant to Ford Motor Company.
Get it together—buckle up.

I've made my living over the years driving cars. Grand Prix cars, sports cars, rally cars, the most honest cars in the world.

You name it, I've driven it.

But in doing so, I've always told a builder just what I thought of his pride and joy.

Blame my skeptical nature.

That's part of it. But the rest of it is the fact that to earn my pay I had to drive the car.

And, at some point, drive it very hard.

And that's what the people at Ford have invited me to do. Drive hard, and tell them what I think...good or bad.

Apparently, they believe the unbiased opinion of a qualified outsider will help build better cars.

They want me to be an honest critic.

It's an unusual position.

But at today's Ford Motor Company, it's not business as usual.

I've seen proof of this in the 1983 Ford, Lincoln, and Mercury cars.

They're full of new ideas.

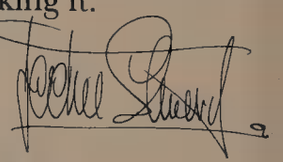
Microprocessors, electronic engine controls, space-age materials, robotic assembly, employee commitment to quality, are just a few of the things I've had to take in while learning about how things are done today at Ford.

Things have changed.

I'm convinced things have changed at Ford. And I'm glad to be part of it. Even though my part may be that of a critic.

To my way of thinking that's all to the good. Most car builders will listen to constructive criticism. But I know one company that goes even further, by actively seeking it.

Today, having a resident car critic is certainly a new way of doing business.



Ford has it. Now.

Ford

FORD · MERCURY · LINCOLN · FORD TRUCKS

Established 1845 SCIENTIFIC AMERICAN February 1983 Volume 248 Number 2

The Future of American Agriculture

In the coming decades the price of farm products in the U.S. will be influenced not only by constraints on the supply of essential inputs such as land, water and energy but also by the demand for exports

by Sandra S. Batie and Robert G. Healy

Agriculture has long been one of the most successful sectors of the U.S. economy. The decline in the proportion of the labor force engaged in farming, a trend attributable largely to the increasing mechanization of agricultural work, has culminated in an exceptionally high level of farm-labor productivity. A mere 3.1 percent of American workers produce not only an abundance of comparatively low-cost food and fiber for domestic consumption but also a substantial (and in some ways troublesome) surplus. Over the past three years U.S. exports of agricultural commodities have averaged more than $40 billion annually, up from less than $10 billion a decade ago. The foreign exchange generated by these exports has been particularly valuable in cushioning the economy against external shocks; for example, since 1973 payments for agricultural exports have offset more than 62 percent of the country's expenditure for imported oil.

In spite of this impressive record, concerns have been expressed about the ability of American agriculture to maintain, let alone expand, its output at reasonably constant prices. For one thing, it is argued, the U.S. no longer has the excess capacity it once enjoyed in such basic agricultural resources as land, water and energy. For another, chronic problems such as soil erosion appear likely to reduce fertility in some areas. There is no guarantee that it will be possible in the future, as it was in the past, to compensate for an ensuing loss of productivity by the genetic improvement of crops, the application of more fertilizer or other technological remedies. Furthermore, disproportionate increases in the price of certain agricultural inputs, particularly energy, will tend to reduce the profitability of many current agricultural practices. These developments, when combined with other factors, such as the widespread adoption of monoculture (the raising of a single type of crop) or the possibility of a major change in climate, may make American agriculture increasingly vulnerable to disruption.

As economists with a special interest in this topic we took part in a recent interdisciplinary study, sponsored by the Federal Emergency Management Agency and the Conservation Foundation, of the factors likely to affect the long-term productivity of American agriculture. Although there was by no means a consensus among the experts, many shared an outlook we would characterize as guarded optimism. We shall explain the grounds for this judgment in what follows, as we review the factors that enable the U.S. to produce food abundantly and cheaply: the quantity and quality of the basic agricultural inputs, the pace of technological innovation and the advantage of a favorable climate.

In a sense there is nothing new about the present apprehension concerning the long-term adequacy of the food supply. Throughout history the question of whether there would be enough food to feed the growing world population has been raised repeatedly. In 1798 Thomas Malthus advanced his famous thesis that the world population was increasing faster than the means of subsistence. Only war, famine, disease or some form of "moral restraint," Malthus argued, could bring the population into equilibrium with the food supply, and then only at a bare subsistence level.

Although there are still millions of malnourished people, for the world as a whole the Malthusian trap has been avoided. The apparent limits to the expansion of agriculture have been circumvented by an assortment of technological stratagems: mechanization, improved plant and animal species, more fertilizer, better control of pests and disease and the development of new cropland through irrigation, deforestation and drainage. As a result higher productivity has been obtained from scarce resources, and comparatively abundant resources have been substituted for scarce ones. In the U.S., for instance, the relative scarcity of labor stimulated first the exploitation of more-fertile cropland and later the introduction of farm machinery. In the early 1800's, before the widespread mechanization of agricultural work, it took more than 250 hours of labor and five acres of cropland to produce 100 bushels of wheat. Today it takes only three or four hours and less than three acres to produce the same amount.

The first major concern about the future of American agriculture we shall address here has to do with the amount of arable land. Considerable attention has been paid recently to the conversion of agricultural land to other purposes, particularly in the course of urban development. According to the best available estimate, made by Michael Brewer and Robert Boxley for the Federally sponsored National Agricultural Lands Study, some 875,000 acres of actual or potential cropland were converted to urban uses annually between 1967 and 1975. This corresponds to an annual conversion rate of .16 percent of the country's total inventory of about 540 million acres of arable land. If the conversion continues at the same rate, 21.9 million acres will be removed from the agricultural inventory by the year 2000.

Is the rate too high? Is the future food supply of the nation therefore in jeopar-

dy? In exploring questions of this kind it is helpful to keep in mind that the land actually used for crops in 1981 was estimated by the Department of Agriculture to be 391 million acres. This exceeds the previous high of 387 million acres, reached in 1949; a modern low of 331 million acres was reached in 1962.

Projections of future cropland requirements are generally based on a worst-case analysis: they begin with a conservative estimate of the prospects for improved yields and a liberal estimate of the growth of food demand. In making both estimates it is assumed that food prices will not increase in relation to other prices. (Rising food prices would limit the need for cropland not so much by reducing caloric consumption as by inducing consumers to reduce their consumption of meat; much of the cropland that is now devoted to growing feed for cattle, hogs and poultry could then be converted to more efficient forms of food production.)

Estimates of the additional cropland that will be needed by the year 2000 range from 26 to 113 million acres. Assuming that potential cropland can be brought into production without undue economic or environmental costs and that some of the land now used for pasture can be converted to crops, there is clearly enough available new cropland (roughly 150 million acres) to rule out a shortage of arable land by the turn of the century, even if the conversion of agricultural land to other purposes reaches a total of 22 million acres by that time. The conclusion remains valid even if one accepts the highest estimates of the need for additional cropland, estimates based on a high demand for agricultural exports, comparatively low yields and no relative price increases for agricultural products.

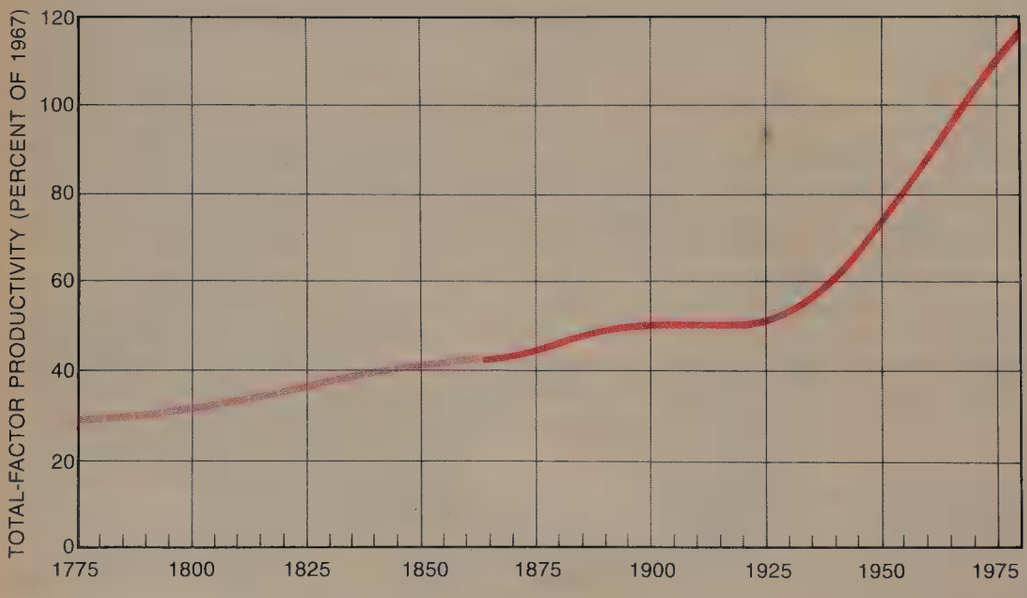
The foregoing analysis does not take into account the fact that much of what is considered potential cropland is already devoted to raising trees and livestock. The conversion of this land to cropland could entail a significant reduction in the supply of forest products and meat. Little research has been done on such possibly important side effects.

Moreover, aggregate statistics on U.S. cropland can be misleading, since they may mask important regional changes in land use. Between 1939 and 1978 roughly a third of the cropland in the Northeastern and Southeastern states was converted to other uses, whereas all other parts of the country gained cropland. Much of the cropland removed from cultivation in the East has gone not to houses, shopping centers and the like but to forests. According to Philip M. Raup, an agricultural economist and land-use expert at the University of Minnesota, "industry has been the chief competitor for farmland, measured not in acres used but in labor force withdrawn. The associated urbanization has generated demands for open space, for recreation and for residential land uses that are inextricably combined with the largely unplanned expansion of private, noncommercial forest land. Our most urban and industrial regions have become the most heavily forested. In the competition for cropland the message to date is clear: local trees have been preferred over local food." Furthermore, in the Southeast much of the former cropland is now pasture, as the South has become an important source of cattle for the feedlots of the Middle West.

Another major competitor for cropland has been the Federal highway program. Much of the interstate highway system was built in the 20 years from 1956 through 1975. It would be difficult to overemphasize the effect of such a vast construction effort on the nation's land-use patterns. The new highways not only took almost two million acres for right-of-way but also brought the possibility of urban and industrial development to additional millions of acres along the interstate routes. Most of the available data on changing patterns of land use in the U.S. were gathered during the period when the effects of the highway-construction boom were at a peak. These statistics form a poor base for projecting future cropland-conversion rates, since the U.S. is unlikely ever again to embark on such an ambitious expansion of the national highway network.

At about the same time that the highways were being built a massive dam-construction program was also under way. The reservoirs flooded large areas of cropland and made much of the adjacent acreage useless for farming. In the Missouri River Basin, for example, low-lying grassland now under water once supplied the winter feed that made possible the summer grazing of the drier, upland areas. As with highways, however, it seems likely that the dam-building era is for the most part ended.

Unlike the effects of highway and dam construction, the full impact of urbanization has probably not yet been felt. Indeed, the distinctions between the traditional patterns of land use are becoming blurred. New mixed-use patterns blend farms and rural residential zones, as migrants from urban and suburban areas repopulate the countryside. Most rural residents are no longer farmers; people living on farms now account for only 10 percent of the rural population. The frictions that arise in such a heterogeneous population may further reduce the profitability of farming and take additional land out of production, beyond what would otherwise be converted to nonagricultural uses.



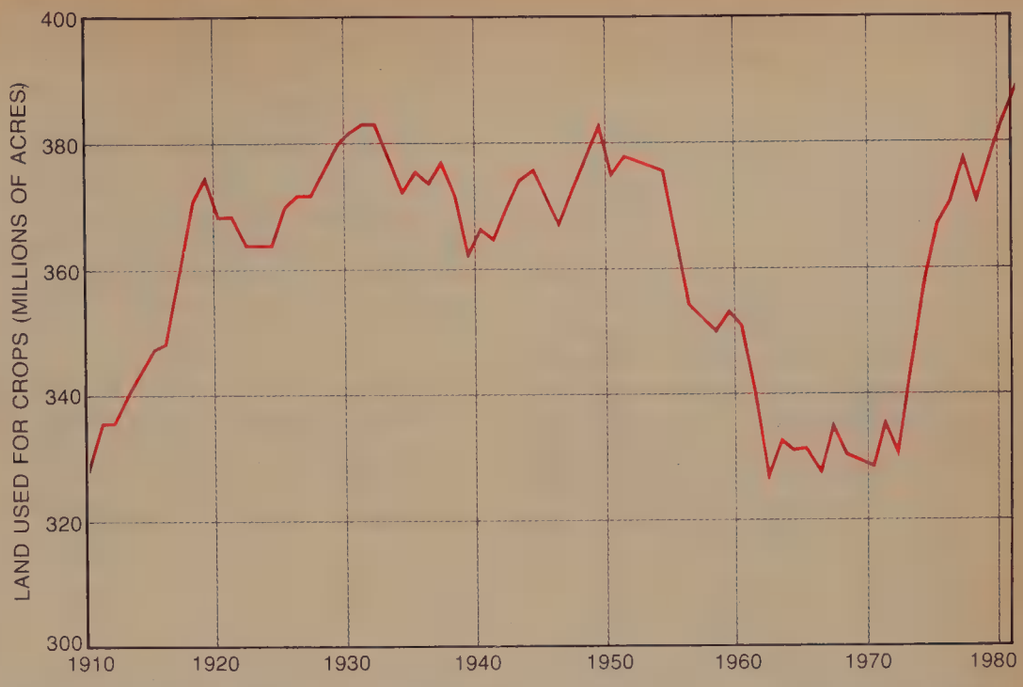
EXTRAORDINARY GROWTH of agricultural productivity in the U.S. over the past two centuries is traced in relative terms. Total-factor productivity is calculated from estimates of the major inputs and outputs of farm production. Data are from Department of Agriculture.

The demand for residential development of rural areas promises to grow as the members of the "baby boom" generation of the 1950's and 1960's become home buyers in the 1980's and 1990's. Some 40 million Americans will reach the age of 25 between 1980 and 1990; the resulting demand for housing may be greater than the U.S. has registered in any previous 10-year period. Land required for recreational purposes, vacation cottages and retirement homes is also expected to increase. Probably as important as direct competition for present cropland is the impact of these residential and recreational demands on the supply of potential cropland. The conversion of such land to nonfarm uses is only part of the process; there are also subtler effects resulting from the division of the land into parcels of uneconomic size or the interspersing of developments that would inhibit the ex-

tension of farming to parcels adjoining present-day farms.

Cropland is also being claimed increasingly for the production of energy. This category of land use includes everything from coal mining to the production of fuels from crops or crop residues. The strip-mining of coal is clearly on the increase, particularly in the Great Plains states, where coal seams accessible from the surface are up to 100 feet thick. It is estimated that coal-mining operations will occupy an annual average of 568,000 acres in the U.S. between now and the end of the century. Of course many of these acres are scheduled for restoration. Although restoration takes time, is expensive and may not perfectly replicate the previous condition of the land, the loss of food production does not appear to be substantial. On the other hand, there is some concern that coal mining and the associated power plants and synthetic-fuel operations could be a major competitor for water in the Great Plains states, thereby creating a new constraint on food production in the region.

More serious in terms of the country's capacity to produce food is the proposed cultivation of crops for fuels such as gasohol. It seems doubtful at present that alcohol made from grain could compete with other fuel sources unless its production were subsidized. Moreover, achieving a 10-percent-alcohol blend in the more than 100 billion gallons of gasoline currently burned per year in the U.S. would require 60 percent of the nation's corn crop. The obvious repercussions on food prices are sobering. The Reagan Administration has so far been reluctant to provide subsidies for biomass conversion, however, and so this fuel alternative is not likely to become a serious competitor for cropland in the near future.



CROPLAND IN ACTUAL PRODUCTION in the U.S. reached a new high of 391 million acres in 1981, up sharply from a modern low period that ended about a decade ago. Potential cropland, including land now in pasture that could be converted into cropland, amounts to another 150 million acres or so, giving the country a total inventory of about 540 million acres of arable land. Estimates of acreage were made by workers at the Department of Agriculture.

In addition to cropland two other basic resources in food production are water and energy. In much of the western U.S. irrigation is essential for high crop yields. In more humid areas irrigation can improve yield, facilitate double-cropping and enable farmers to plant more valuable crops. In the country as a whole irrigated acreage has tripled since 1940, and it now accounts for nearly a fourth of the value of the nation's crops. The large increase in irrigated land can be traced to the fact that in many areas water was long considered a free commodity, available to the farmer for no more than the price of transporting it. Low energy prices and Federal subsidies kept the cost of transportation low.

The era of low-cost water is past. In areas where farmers rely on ground water a falling water table and increased pumping costs are making the recovery of the water more expensive. High energy costs also make it less attractive to pump surface water over long distances. In the West most surface water is already fully allocated, and so new consumers will increasingly have to bid water rights away from present consumers.

Much of the recent growth in irrigated acreage has been concentrated in four states: Nebraska, Kansas, Oklahoma and Texas. Here irrigation is used mainly to grow sorghum, corn and alfalfa. These crops are commonly raised to feed cattle, and many of the nation's largest feedlots are in the same states. The beef production has been made possible by the "mining" of the Ogallala aquifer, a vast underground lake that spans eight states [*see illustration on next page*]. Recharge to the aquifer is exceedingly slow, with the result that the water table has been steadily falling and pumping has become more costly.

Eventually the combination of a falling water table and rising energy costs will probably make ground-water irrigation less popular. Farmers are adjusting to higher water costs by adopting water-conservation measures; still, the difference in profitability between irrigated farming and dryland farming has narrowed. If dryland farming returns to these areas, concentrated cattle feeding will no longer have a competitive advantage in the states overlying the Ogallala aquifer. As Raup has stated, "unpriced water . . . has been capitalized into a national level of beef consumption that cannot be sustained in the long run without a return to the feed-grain supplies of the Corn Belt. We have a fed-beef economy that has become dangerously dependent on an exhaustible resource base."

Whatever happens to the feedlots, it is clear that unless grain prices increase to offset the increasing cost of water, irrigation cannot continue indefinitely to contribute to agricultural expansion as it has in the past three or four decades. Kenneth D. Frederick of Resources for the Future points out that this is not necessarily a bad thing. He comments: "If the transition from water abundance to scarcity allows for an efficient use of resources over time, irrigation will contribute to agricultural production and growth for many more decades. The serious problem will emerge if we attempt to keep water cheap when it is not. Such a policy will ensure its inefficient use . . . thereby limiting the long-term role of irrigation."

Energy is consumed on the farm for many purposes. Irrigation has already been mentioned, but agricultural chemicals and vehicular fuels make even larger demands. The historical trend has clearly been toward greater energy consumption. Herbicides are now applied to more than 90 percent of the acreage planted in corn, whereas a decade ago only 57 percent of the corn acreage was treated with such chemicals. Similarly, the horsepower of tractors increased by 31 percent between 1966 and 1978.

The economic reasons for this trend are not hard to find. Between 1950 and 1978 labor costs on the farm went up twice as fast as energy prices, and land prices rose almost four times as fast. The incentive was to substitute energy—in the form of chemicals, fuel and fertilizers—for both land and labor. The price

relations were accentuated by government policies that tended to keep the prices of fuel and fertilizer as low as possible. Natural gas, for example, has been kept at an artificially low price by government regulation, and electricity is supplied to most farms on a highly favorable basis.

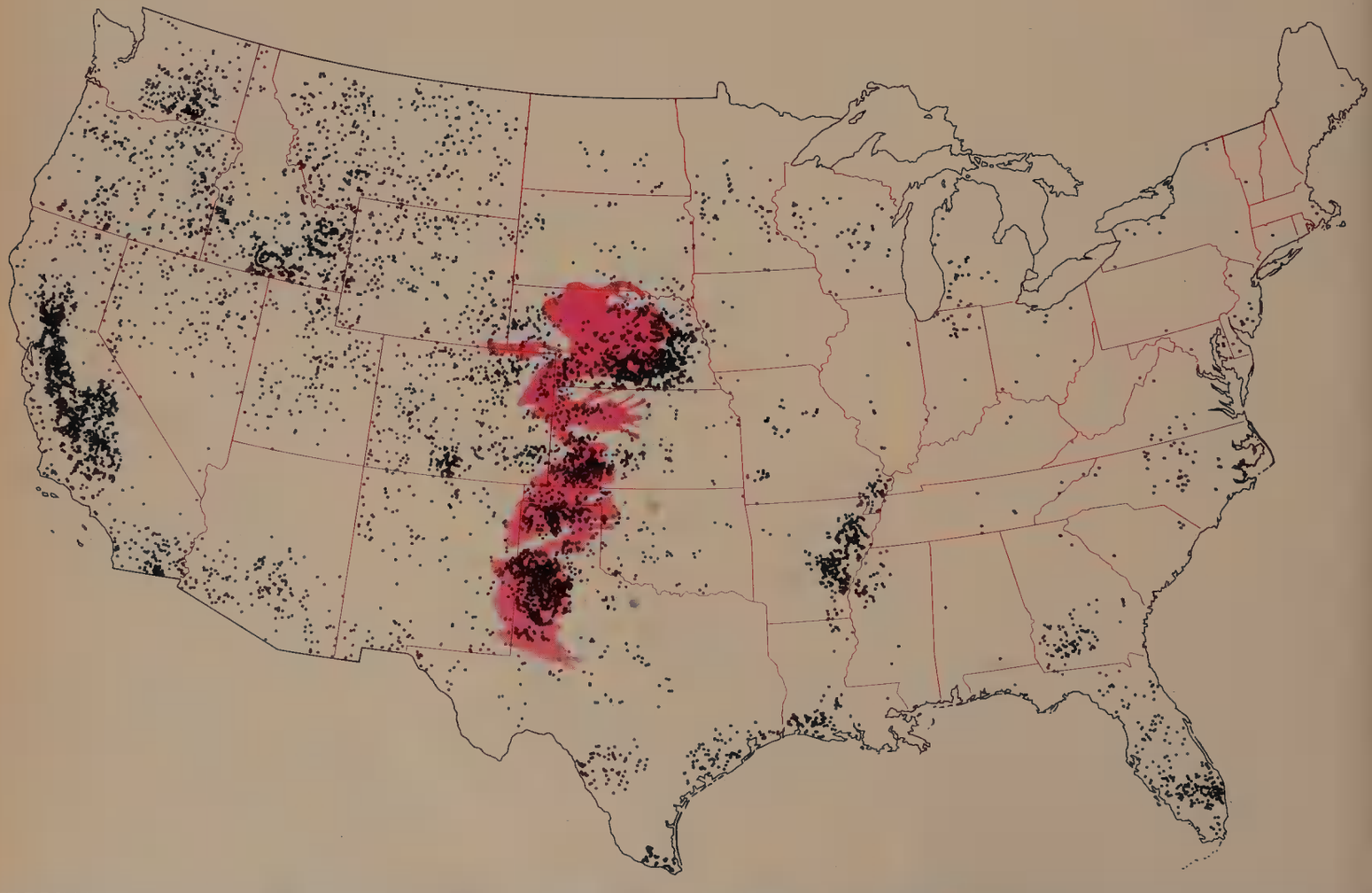
Low-cost energy has been exploited in agricultural production not only to reduce the need for farm labor but also to increase the productivity of all other inputs and to reduce the risk of crop failure or spoilage. The reduction of risk can take many forms: limiting the variability of the harvest through the application of pesticides, reducing the adverse effects of climate through irrigation and reducing the risk of spoilage by drying or refrigeration.

Because energy performs so many vital functions for the farmer, agricultural energy consumption is unlikely to be reduced by much, in spite of rising energy prices. Relative energy prices do not appear to be increasing fast enough to disrupt production or to exert much pressure on food prices. There may be some shifts in the location of farm enterprises with rising energy prices, however. Higher transportation costs may make it feasible to fatten cattle on feedlots in the South rather than shipping them to the Middle West. Local crops such as tomatoes may be better able to compete with nonlocal crops that are cheaper to produce but more expensive to transport. Technological and institutional changes are also likely to facilitate energy conservation; better irrigation scheduling and integrated pest-management planning are two examples. Some farms, such as dairies, may even resort to meeting part of their own energy demands through the recycling of wastes.

The bounty produced by American farmers currently accounts for 13 percent of the world's wheat, 29 percent of the coarse grains (mainly corn), 17 percent of the meat and 62 percent of the soybeans. Because the U.S. has less than 5 percent of the world population, the high level of domestic food consumption still leaves a large surplus for export. Nevertheless, there is concern that further increases in output may not be sustainable unless the quality of the land is maintained. The chief problems are the erosion of soil by wind or water, the compaction of soil by heavy farm equipment, the increasing salinity of irrigated soil and certain side effects of schemes for the production of energy from crop residues.

Erosion can remove large amounts of topsoil and still not be obvious to the casual observer. A farm losing 40 tons of soil per acre per year loses less than a third of an inch of soil per year, which is not a readily visible amount. Water-caused soil erosion is most serious in the Southeastern states, where a single rainstorm can sweep away as much as an inch of topsoil from a cleared field. Significant erosion, however, is recorded on only a small fraction of the nation's cropland. Water-caused erosion associated with soybean production, for example, is concentrated in several major agricultural regions in the South and Middle West. Serious wind erosion is observed mainly in the High Plains region of Texas, southeastern Colorado and southwestern Kansas, where wheat is a major crop [*see illustration on pages 50 and 51*].

Are soil-erosion rates today greater than they were in the 1930's? In spite of many statements to this effect the question is probably unanswerable. There is some evidence, however, to suggest a recent increase in soil erosion. John F. Timmons of Iowa State University has shown that in Iowa annual soil losses decreased from 21.1 tons per acre in 1949 to 14.1 tons in 1957, but the rate rose again to 17.2 tons per acre in 1974. The reversal of the trend can be ascribed for the most part to the rapid rise in the prices of wheat, corn and soybeans in the early 1970's, a development that



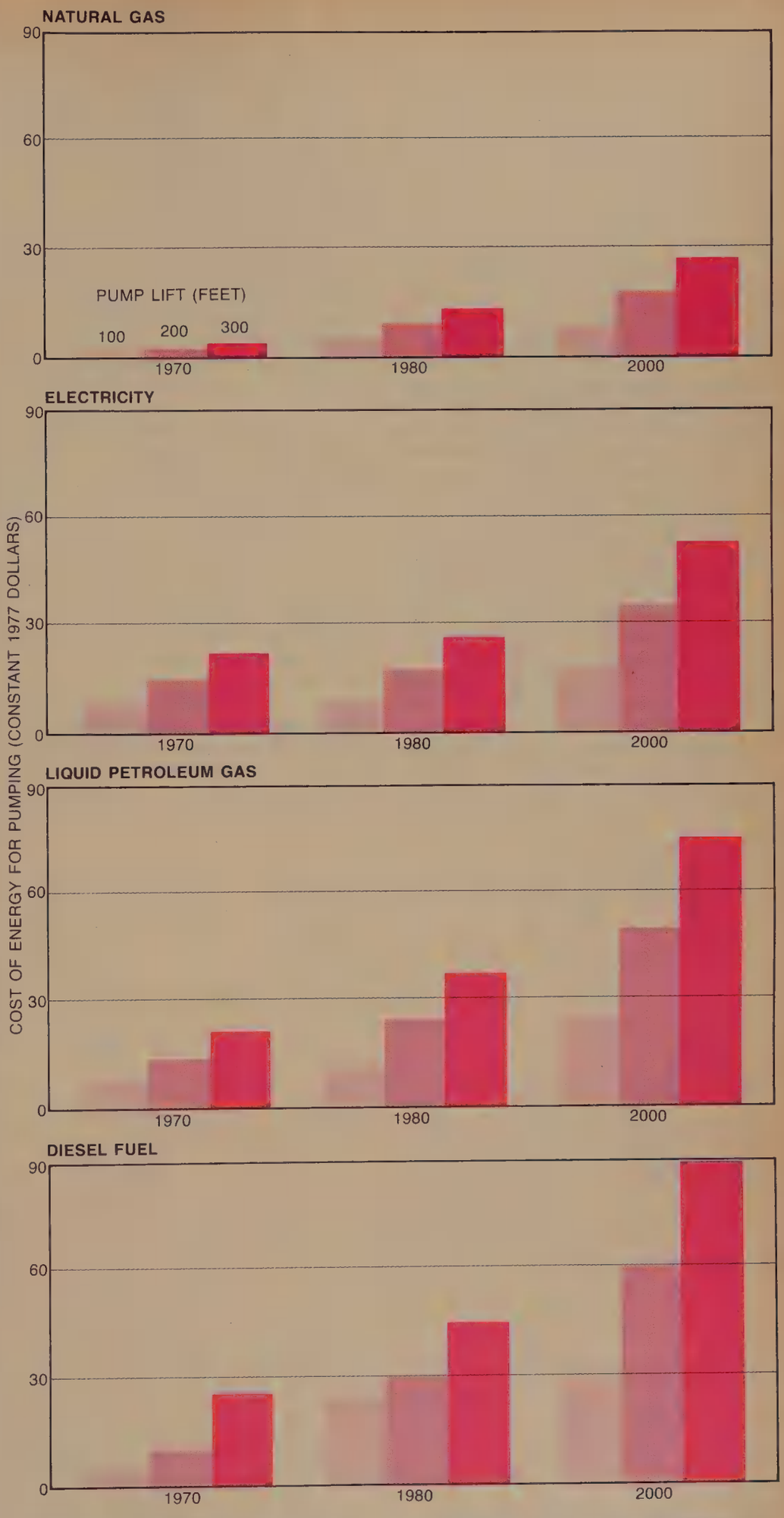
IRRIGATED LAND IN THE U.S. totals about 60 million acres. Each dot on the map represents 8,000 acres where irrigation facilities existed in 1977, the latest year for which information of this kind is available. The colored area shows the lands underlain by the Ogallala aquifer, where much of the postwar growth in irrigated acreage has been concentrated. Crops are raised here mainly to feed cattle.

caused erosion-prone land formerly in pasture to be plowed for crops. Land devoted to these crops has increased further since 1974.

Although an increase in the acreage devoted to the production of corn, soybeans, wheat and cotton in response to a projected increase in the demand for food and fiber could lead to more erosion, the advent of "conservation tillage" may lessen the effect. Conservation tillage is defined as any method of preparing the soil that reduces the loss of soil in water runoff compared with conventional unridged, or clean, tillage. Common conservation-tillage practices include no-till, minimum-till, ridge-plant, plow-plant and chisel-plowing techniques. Since the 1960's conservation tillage has spread rapidly, and it is now practiced on an estimated 25 percent of the tilled cropland. The rapid growth of conservation tillage has come largely in response to economic incentives: it generally costs less than conventional tillage in both labor and fuel. Still, the result is also less erosion. Pierre R. Crosson of Resources for the Future predicts that by the year 2010 between 50 and 60 percent of the nation's cropland will be farmed by means of conservation tillage. He also estimates, however, that productive cropland may increase by as much as 75 million acres by 2010 and that "much of that additional land would be more erosive than land now in crop production." Thus substantial amounts of erosion may still prevail.

Soil erosion can affect productivity by reducing the capacity of the soil to hold water, by removing plant nutrients, by degrading the structure of the soil and by reducing the uniformity of the soil within a given field. The severity of these effects differs widely, depending on the region, the soil, the farming practices employed, the crop and the length of the growing season. Assessing the reduction in yield caused by erosion at a specific site is difficult because many other factors can affect yield. As erosion has progressed over the years, farmers have applied more fertilizer both to substitute for eroded topsoil and to enhance fertility. Farmers have also changed tillage practices, introduced different types of machinery and adopted new varieties of plants.

In spite of these complicating factors, experts who have studied the effect of erosion on productivity conclude that in some cases the loss of even an inch of topsoil may be significant. Where erosion has substantially reduced the topsoil depth, so that plant roots are reaching less fertile subsoil, the farmer's costs of production are higher because more fertilizer is needed to maintain yield. Furthermore, adding nutrients in the form of fertilizer can increase yield only if enough water is available in the soil;



ENERGY COSTS of lifting an acre-foot of water from various depths are estimated in this bar chart for the years 1970, 1980 and 2000. Projected increases in pumping depth, brought on by the drawdown of aquifers, combined with rising fuel prices, brought on in part by deregulation, are likely to result in a sharp increase in the costs of pumping, a development that is expected to reduce the competitive advantage of irrigated farming over dryland farming in many areas. The chart is based on a study by Kenneth D. Frederick of Resources for the Future.

hence differences in the capacity of soils to supply moisture are often the primary determinants of differences in crop yield.

In addition to reducing productivity soil erosion is also the main contributor to agricultural water pollution. Runoff from farmland carries sediment, nutrients, herbicides and pesticides, all of which are considered pollutants when they are present in excess. Sediment washed off farms is reducing the lifetime of many inland lakes and reservoirs and increasing the cost of maintaining navigable channels in rivers. Excessive nutrients can lead to the eutrophication of bodies of water and can raise the cost of purifying public water supplies. Excessive herbicides and pesticides can be detrimental to wildlife. Because the damage happens off the farm, however, farmers have little or no incentive to take such factors into account.

Technical solutions to the problem of soil erosion exist. Steep lands can be terraced. Crop residues can be left in place. Conservation tillage and contour plowing can be practiced. Runoff can be collected in basins. In the case of conservation tillage most studies show a reduction in erosion of between 50 and 90 percent, depending on the percentage of the field left covered by crop residues.

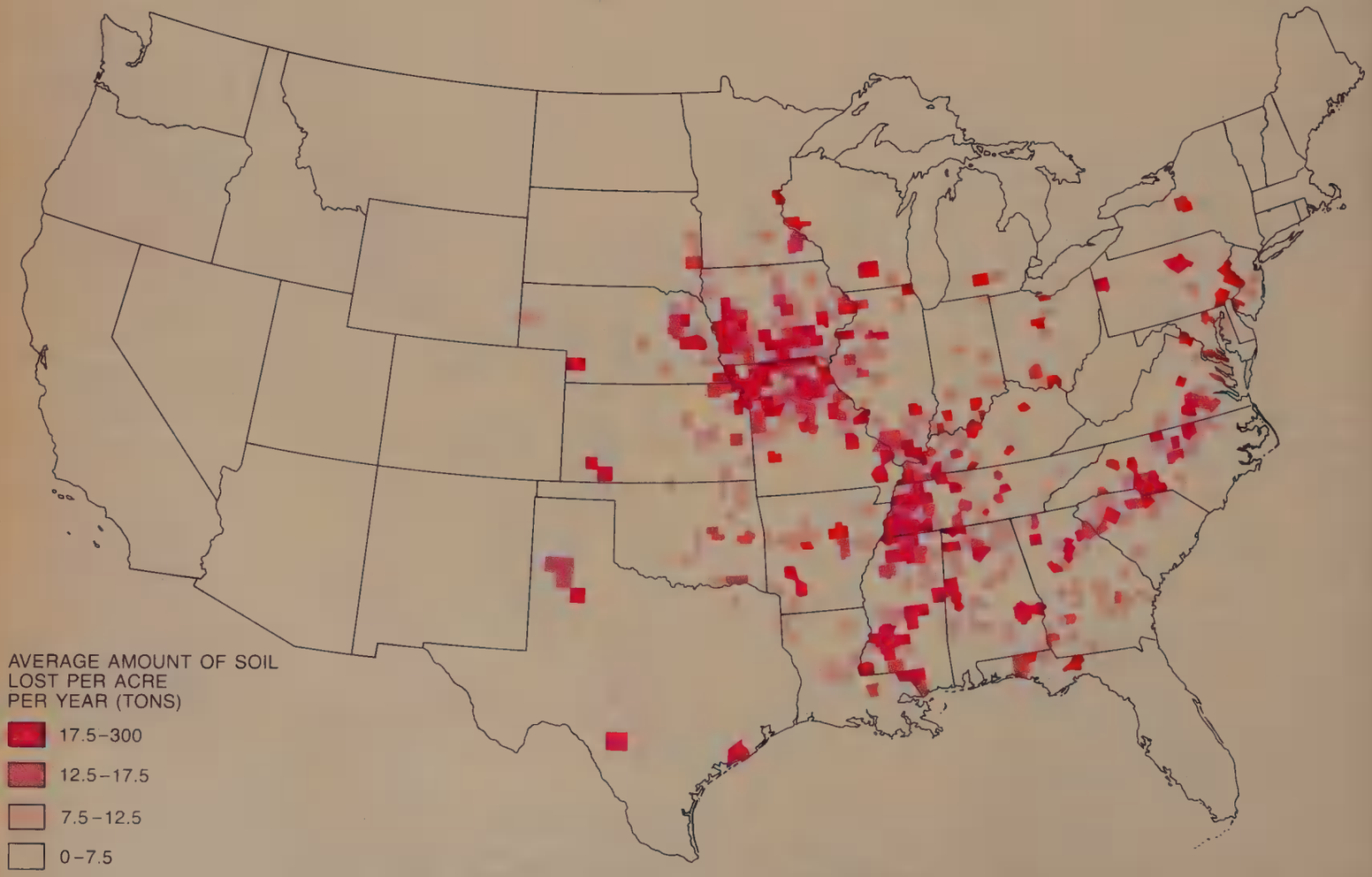
If such measures can control erosion, why are they not more widely practiced? Contributing factors may include the personal preferences and beliefs of farmers, the physical characteristics of the land, tenure arrangements, tax policies and farm-commodity programs. The main reason, however, is simply that many conservation practices do not pay for themselves. Because erosion degrades productivity and water quality, because there are few incentives for the farmer to adopt most of the conservation practices and because erosion is likely to increase as the nation responds to the demand for exports, many observers have reached the conclusion that erosion is a problem worthy of greater public concern and management.

Heavy farm machinery can lower productivity by compacting the soil. Compaction can reduce water percolation, resulting in drainage problems and increased runoff, and it can inhibit root penetration. Some plants, such as potatoes and sugar beets, are more sensitive to these effects than other plants. Compacted soils can also inhibit the transmission of gases and heat, another effect detrimental to the growth of most plants.

Instances of lower crop yields caused by soil compaction have been reported, but there is also some evidence to the contrary. For example, soybean yields in Minnesota have been shown to be higher on moderately compacted soils than on noncompacted soils. Actually little is known about the extent or significance of soil compaction. In any case, technologies and management practices are available to relieve problems caused by soil compaction. They include careful timing of farming operations, reducing the number of trips into the field, confining wheeled traffic to designated paths, deeper plowing and wider equipment. A recent study by the Congressional Office of Technology Assessment concluded that existing technology can greatly reduce the possibility that soil compaction will become an important constraint on productivity.

Excessive salinity is considered the most pervasive environmental problem associated with irrigated agriculture. When irrigation water is applied to crops, evaporation and transpiration remove the pure water, leaving salt as the residue. If rainfall or further irrigation does not flush the salt away, the productivity of the soil can suffer. Between 25 and 35 percent of the irrigated western croplands are thought to have excessive salinity.

Two areas with serious salinity problems are the Lower Colorado River Basin and the west side of the San Joaquin valley of California. In the San Joaquin area it is estimated that yields have already been reduced on more



SOIL EROSION is a serious problem on only a small fraction of the nation's cropland, although the areas affected include some of major importance for food production. The maps on these two pages show the average tonnage of soil lost per acre to water- and wind-caused

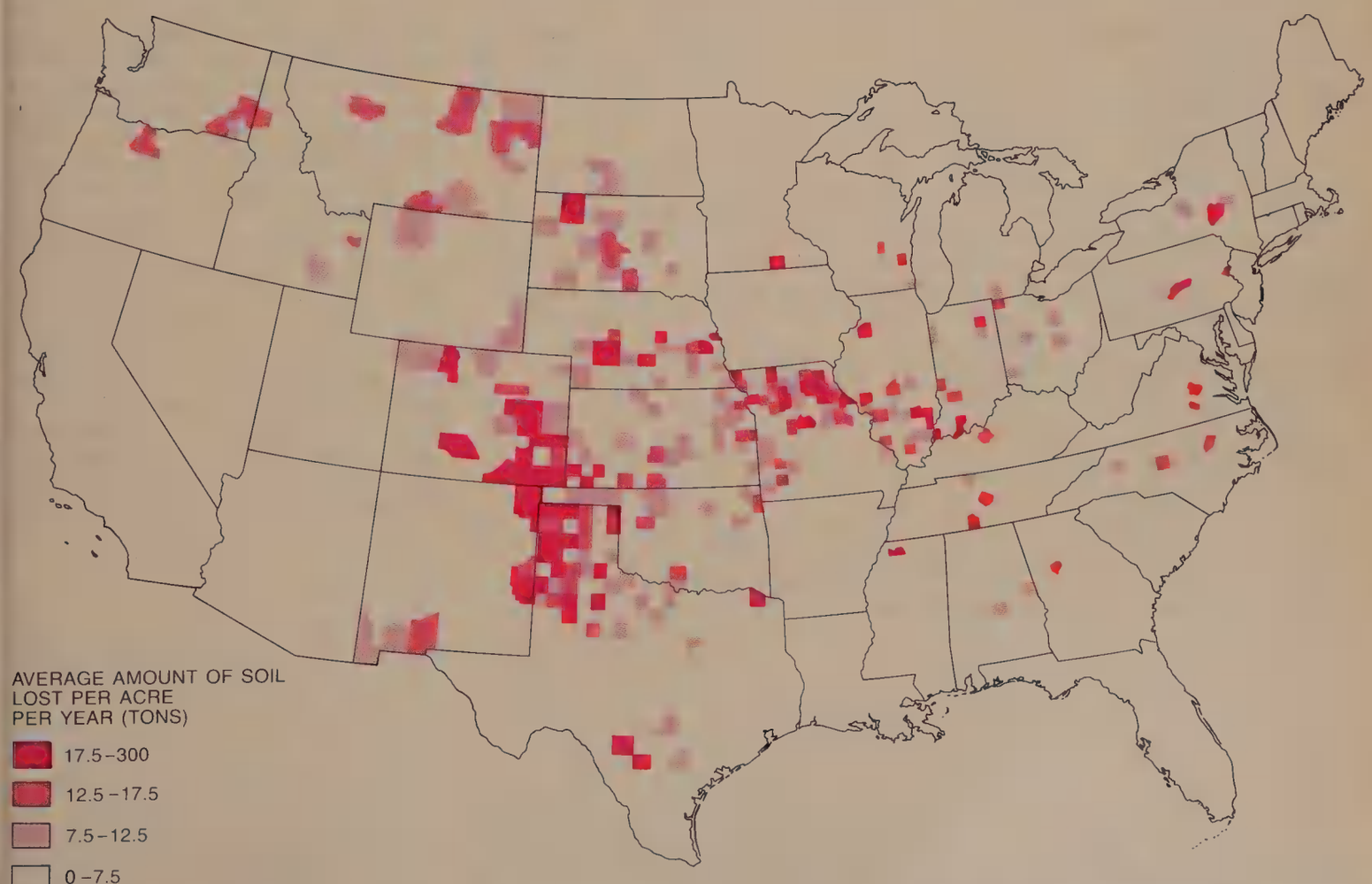
than 400,000 acres, with an estimated economic loss of $31.2 million per year. Farmers have asked for government investment to provide water of higher quality, but in the meantime there are other remedies farmers could adopt themselves. One technique is to apply irrigation water more efficiently through the use of drip-irrigation systems or lined irrigation ditches. Another approach is better planning for the recirculation of the irrigation water. The first use of the water would be for those crops most sensitive to salt. Subsequent uses (as the water became saltier) would be on less sensitive crops. Regardless of who pays for the investments needed to reduce the salinity of soils, it appears that at least some investment would be worthwhile.

The rising cost of energy has led many people to consider alternative energy sources. One possible alternative is to exploit crop residues for direct burning or for conversion into fuel. Although these schemes are technically feasible and enormous quantities of residue are available (an estimated 400 million dry tons annually), such an approach is not necessarily either economic or desirable. Crop residues are bulky and not easily transported, and they would have to be hauled at considerable cost to the plants where they would ultimately be converted. Several studies have found that various crop residues would be at an economic disadvantage with respect to coal when prices are calculated at the point of delivery to local utilities. Furthermore, residues left in place have a useful role in reducing soil erosion and recycling nutrients. For this reason farmers in many areas are reluctant to have crop residues removed. In short, the potential of crop residues as an alternative energy source is constrained by both economic and environmental considerations and in any case is nowhere near the 400 million dry tons available in principle. One estimate is that at most there would be 80 million dry tons available annually.

Granted that the quantity and the quality of agricultural land and other inputs are important to the future of American agriculture, it is nonetheless true that the Malthusian specter has been avoided so far mainly because of technological advances that have greatly improved productivity for a given level of inputs. Here too there are concerns that the growth of productivity cannot be sustained at the high level of the past few decades. Even some of the more optimistic investigators think the U.S. will do well simply to maintain the same rate of absolute yield increases in the future.

The problem is not really a lack of possible technological breakthroughs. There are plenty of these in prospect: the hybridization of crops such as wheat, the enhancement of photosynthetic efficiency, the twinning of livestock, the breeding of drought-resistant plants and so on. Instead many observers fear that the development of these technologies will be hampered by limited expenditures on research and development and on the extension programs that bring the results of such research to the farm. There has been little growth in Federal support for agricultural research since 1975, compared with the 3.8 percent annual growth rate from 1939 through 1965. Private-sector expenditures for agricultural research and development are difficult to estimate, but they are probably as large as the public-sector expenditures. Even if the private-sector research increases, however, the expenditures are more likely to be for innovations that are patentable, such as machines, drugs and pesticides.

The projected rate of technological change is of crucial importance; it has been technological progress that has enabled the U.S. to meet food demands at comparatively low prices without being faced with severe resource constraints. If increased productivity is not forthcoming or is achieved only after a long delay, and if predicted future demands do indeed materialize, then either prices



erosion in 1977. The map at the left shows the erosion of cropland used to raise soybeans; most of the erosion in this case is attributable to water. The map at the right shows the predominantly wind-caused erosion of cropland used to raise wheat. Keys indicate erosion rates.

will have to rise, exports will have to be curtailed or substantially more resources will have to be diverted from other needs and devoted to food production.

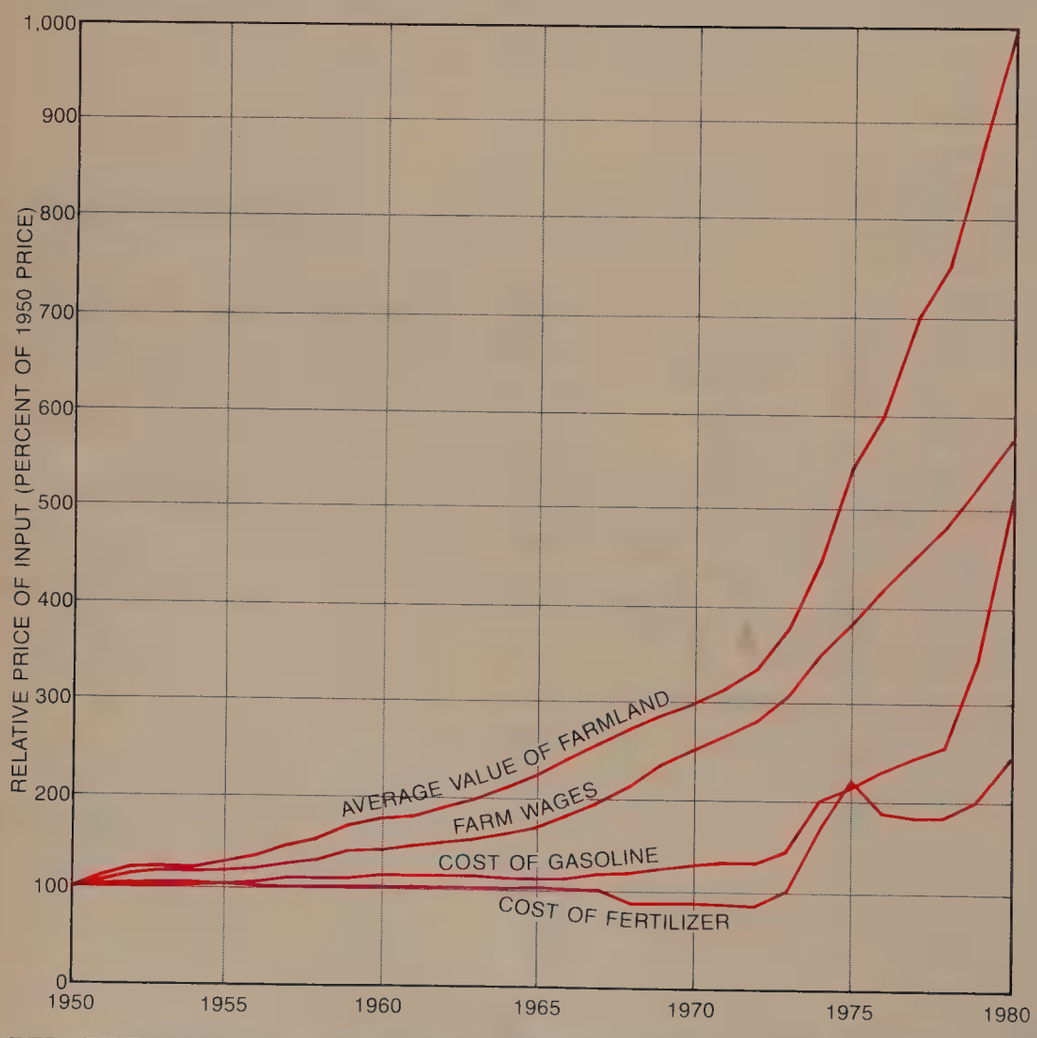
The technological revolution that agriculture experienced over the past three decades was accompanied by enormous changes in what is often called the structure of agriculture. Between 1950 and 1980 the number of people living on farms fell from 23 million to fewer than eight million and the average size of farms doubled, to just over 400 acres. Although a fairly large number of small farms still exist, the bulk of U.S. food production comes from a diminishing number of farms. Workers at the Department of Agriculture have predicted that if present trends continue, by the year 2000 the nation's 50,000 largest farms will account for 63 percent of all agricultural sales, up from 31 percent in 1974. Also projected is increased specialization by individual farms in a limited number of commodities, in contrast to the product diversification and virtual self-sufficiency that once characterized many family farms in the U.S.

The structure issue is important mainly for its political and social ramifications, but it may also have a bearing on future agricultural productivity. Increasing scale has often been accompanied by greater mechanization and more irrigation. Both have made the agricultural system more dependent on energy and other inputs produced off the farm. Crop specialization by the individual farm has led to localized monoculture, whereas regional specialization has made the food-production system more vulnerable to climatic variation and the distribution system more susceptible to the effects of energy shortages and transportation stoppages.

Some of the same technological advances that have made American agriculture so productive have also reduced the ability of some crops to withstand pests and disease. Modern plant-breeding practices have resulted in crops with comparatively little genetic variability. More than 70 percent of the acreage planted in corn, for example, is planted with the descendants of only six inbred varieties. The narrow genetic base of major crops such as corn can make them more vulnerable to disease, insects or erratic climate.

The vulnerability was vividly demonstrated in 1970 when a mutation of the blight-causing fungus *Helminthosporium maydis* caused an epidemic of corn-leaf blight in the U.S. An estimated 15 percent of the corn crop was lost. Although seed companies quickly introduced resistant seed and recovery was rapid, the single-season epidemic was perceived as threatening. Farmers became aware that the stability that comes from ecological diversity had been sacrificed in order to get the increased productivity and reduced production costs associated with plant uniformity.

There are ways of reducing the risks of having monocultured crops. The improvement of germ-plasm collections would make it possible to respond quickly to epidemics or pest infestations. Strengthening plant-breeding programs and upgrading the monitoring of diseases and pests can also improve and protect plant resiliency. Agricultural research will have to be better funded and in some respects reoriented if these and other improvements are to be implemented.



RELATIVE PRICE INCREASES of four major agricultural inputs—land, energy, labor and fertilizer—are plotted here for the period from 1950 through 1981. The largest increase by far has been in the average value of farmland, which has risen by more than 900 percent over the past three decades. The graph is based on data gathered by the Department of Agriculture.

Of all the concerns about the vulnerability of American agriculture the possibility of an adverse change in the climate is probably the most difficult to analyze. The difficulty results less from limited data than from limited understanding of the reasons for climatic variation. The available evidence does suggest that agriculture has developed in an abnormally warm period. Furthermore, warm periods such as the present one appear to last for no longer than 10,000 years, and this one has already lasted for almost that long. Climatologists do not agree on whether the earth is now in a long-term warming trend or a long-term cooling trend. In any case, long-term climatic trends are probably not as important to American agriculture as annual fluctuations in weather.

Human modifications of climate further complicate the analysis. The cutting of forests, the creation or expansion of deserts, the production of acid rain and the addition of particulates, ozone and carbon dioxide to the atmosphere can change climate, but the ultimate impact on crop yields is uncertain. Some crops such as tomatoes and strawberries may actually benefit from more-acidic rain, whereas others such as lettuce and spinach could be injured.

In a report published in 1980 the National Defense University made an effort to predict changes in crop yields under various conditions of temperature and mean precipitation. Agricultural experts were asked to estimate yield changes for several combinations of conditions during the reproduction period of various crops in various countries. The estimates for corn in the U.S. showed that maximum yields would be achieved if the temperature were slightly cooler during the reproduction period and if precipitation were 20 percent

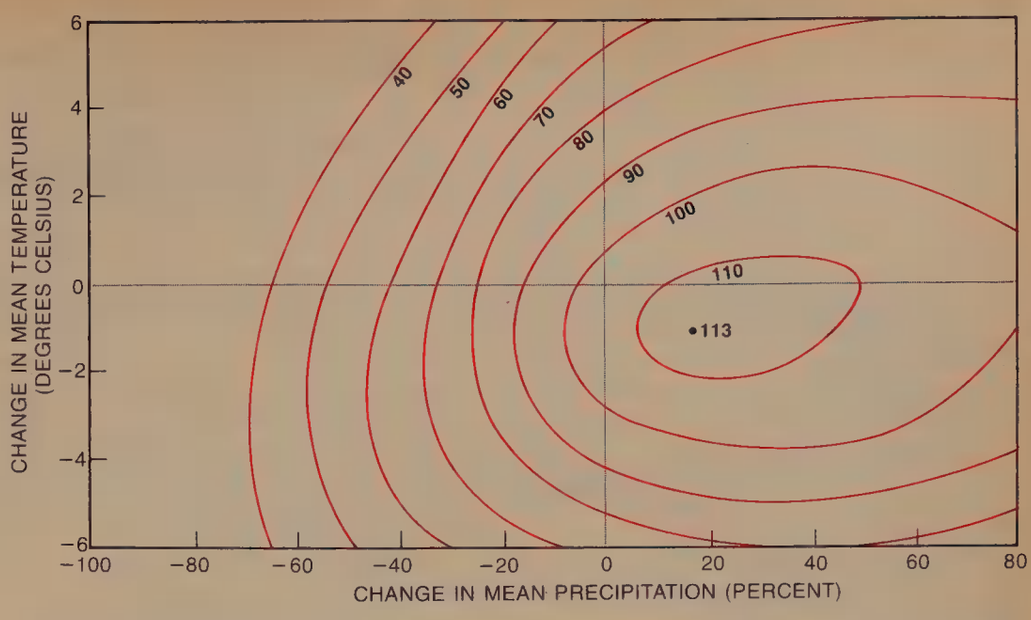
above normal [*see illustration at right*]. Soybeans, on the other hand, would reach a maximum yield with no change in temperature but with a 40 percent increase in precipitation. Although predicting climatic change and its effects remains guesswork, it is safe to assume there will be some unfavorable weather.

There is no simple answer to the question of whether the bounty of American agriculture can be sustained in the long term. Few of the experts, if any, foresee circumstances in which future agricultural output will be insufficient to feed the domestic population, which is now growing at less than 1 percent per year. The real uncertainty lies in the future level of exports, which already claim the output of one out of every three acres of U.S. cropland. So far agricultural exports have been regarded as an unqualified benefit to the U.S., raising farm income, generating foreign exchange and absorbing the chronic grain surpluses that have plagued American farmers at various times, including the present. Until recently the only serious objection raised to farm exports was based on their impact abroad; the argument was that they kept prices in the consuming countries so low that local farmers found it impossible to compete.

As the cushion of excess capacity in U.S. farming has diminished, however, a few voices have begun to ask whether unlimited export expansion is in the national interest. Should we encourage foreign nations to become dependent on U.S. grain at low prices that may not be maintained in the long run? If American agriculture is pushed to its limits, they argue, it will not be by domestic demand alone. The future expansion of farm exports is a fundamental policy question that deserves the same attention that has recently been given to the threats to agricultural supply.

On the supply side several of the most important constraints—those on land, water and energy—will probably manifest themselves not through physical shortages but through rising prices. Farmers' responses to rising input prices may bring increasing pressure for interregional shifts in crop and animal production as well as changes in production practices. Increased irrigation costs and rising demand for grain, for example, may move the livestock industry back to the grazing patterns of the 19th century. If that happens, cattle herds will put greater demands on the land than ever before, and higher prices for both cattle and land will be the likely outcome.

Other concerns—soil erosion, excessive soil salinity, climatic change and lags in technological growth—seem less tractable. If future increases in productivity are more in line with the moderate growth experienced between 1900 and 1950 than with the more recent high-growth era, and if increased export demands for American agricultural products do indeed materialize and are accommodated, then we may witness a return to an agriculture in which increased production depends mainly on the cultivation of more land rather than on more output per acre. The problems of reduced soil fertility and adverse climatic change would then be far more serious than they would be if increased productivity were likely.



EFFECT OF CLIMATIC CHANGE on the estimated yield of the U.S. corn crop was forecast as part of a recent study done by the National Defense University. The results, displayed in the form of a contour diagram, showed that the maximum corn yield would be achieved if the temperature were slightly colder than normal during the crop's reproduction period and if the precipitation were 20 percent above normal. The contours connect combinations of temperature and precipitation that would result in the same relative yield. For example, the curve marked *80* shows the climatic conditions under which yields are projected to be 80 percent of the average yield. Similar diagrams were generated for a number of other crops in the U.S. and abroad.

The high degree of crop monoculture that characterizes American agriculture is a serious threat only if the agricultural establishment fails to take some simple preventive measures, such as maintaining seed collections and a healthy and diversified seed industry. Unfortunately government has been somewhat reluctant to take the necessary leadership role in this area.

In general, predictions based on past and present trends do not suggest anything approaching a crisis in American agriculture. The predictions describe only the most probable events, however, and not those that are less likely but still possible. Perhaps American agricultural policy should be focused more sharply on insuring against extreme deviations from the likeliest event. The most obvious example of such a deviation is a change in the climate to one less favorable to agriculture. The amount of insurance that might be appropriate would depend, of course, on the costs and benefits of the preventive strategy selected. The possibilities include wider spatial distribution of agricultural lands; more diversity in agricultural production by region; increased funding for agricultural research; the establishment of private and public grain reserves; programs to reduce specific threats to agriculture; modifications to government programs that encourage a structure of agriculture vulnerable to extreme events, and adjustments to the prices of agricultural products to reflect more of the private and social costs of food production.

Needed research relevant to these strategies would include work on determining the future costs of soil erosion, compaction and salinity; on gauging the effects on agricultural productivity of a severe climatic change; on assessing the costs of converting potential cropland into actual cropland, and on estimating the trade-offs in productivity resulting from wider and more diverse spatial distribution of agricultural enterprises. In most situations it will be easier to identify the costs incurred when implementing a preventive strategy than it will be to estimate the benefits from protecting against an extreme event.

The debate over American agricultural policy should not await research findings, if for no other reason than that policy debates can help to identify areas of ignorance. With the present farm economy beset by production surpluses and cost and credit squeezes on the farmer, it may seem strange to suggest debating how the U.S. might confront possible long-term shortages. The urgency of short-term problems, however, must not obscure the importance of considering how to organize an agricultural sector that can survive and flourish for as long as we expect our society to endure.

The Lattice Theory of Quark Confinement

The force between quarks in a particle such as the proton has been simulated by imposing a discrete lattice on the structure of space and time. The results suggest why a free quark cannot be isolated

by Claudio Rebbi

The development of quantum mechanics put to rest the uncritical acceptance of the idea that elementary particles are the "building blocks" of matter. Often such particles do not act like hard, impenetrable blocks at all, and in many circumstances they must be described as waves. Until recently, however, it still seemed that elementary particles were like building blocks at least to the extent that each particle could in principle be isolated and observed as an individual entity. The electron, the proton and the neutron, for example, can be separated from one another and observed as individual packets of waves. Even this limited interpretation of the building-block metaphor fails in the case of the quark, the supposed constituent of the proton, the neutron and many related particles. Apparently a quark cannot be isolated; although there is abundant evidence for the existence of quarks and antiquarks bound together in pairs and triplets, an individual, or free, quark has never been observed.

As the experimental evidence has accumulated, it has begun to seem that if quarks are real particles at all, they must be permanently bound within nuclear particles. Any theory of quark interactions ought to account for this phenomenon, which is called quark confinement. It is easy to construct pictorial models of particles such as the proton in which the constituent quarks are confined. For example, the quarks can be thought of as being fastened to the ends of an unbreakable string; they are then free to move about within the volume defined by the length of the string but cannot wander away from one another. It is a formidable task, however, to formulate a theory that can account for the permanent binding of quarks and the structure of nuclear particles without violating the constraints imposed by the theory of relativity, quantum mechanics and the principles of ordinary causality.

After several years of both experimental and theoretical investigations most particle physicists are confident they at last have a theory capable of explaining the interactions of quarks. One reason for confidence is that the theory is a mathematical analogue of the most successful physical theory ever developed: the quantum theory of interactions in an electromagnetic field. The latter theory is called quantum electrodynamics, or QED, and the conceptual similarity of the theory of quark interactions to QED is reflected in the name of the new theory: quantum chromodynamics, or QCD.

The difficulty that has delayed full acceptance of QCD is that its mathematical complexity makes any rigorous, analytical prediction from it exceedingly difficult. Indeed, up to now the most eagerly sought prediction of QCD, namely the demonstration of quark confinement, has not been forthcoming. Recently, however, my colleagues and I at the Brookhaven National Laboratory have applied mathematical methods that rely heavily on the capabilities of the digital computer to the problem of confinement, and a numerical breakthrough has been achieved. Because the method explores the implications of QCD by making a series of increasingly accurate approximations, the results of the calculations do not carry the same force as a logical deduction from accepted first principles. Nevertheless, the numerical results have provided strong evidence for the confinement of quarks.

The framework for the calculations is a pioneering suggestion made in 1974 by Kenneth G. Wilson of Cornell University. Wilson proposed that QCD be formulated on a cubic lattice, an array that divides space and time into discrete points. The lattice is only an approximation to real space-time, but it allows calculations to be made that would otherwise be impossible. As the mesh of the lattice is made progressively finer the values of physical quantities defined on the lattice converge to the values QCD would predict for them in ordinary, continuous space and time. Our numerical approximations show that for an extremely fine lattice, confinement is a consequence of QCD. For reasons that will become clear both QCD and QED are called gauge theories; the computational method I shall describe is therefore called a lattice gauge theory.

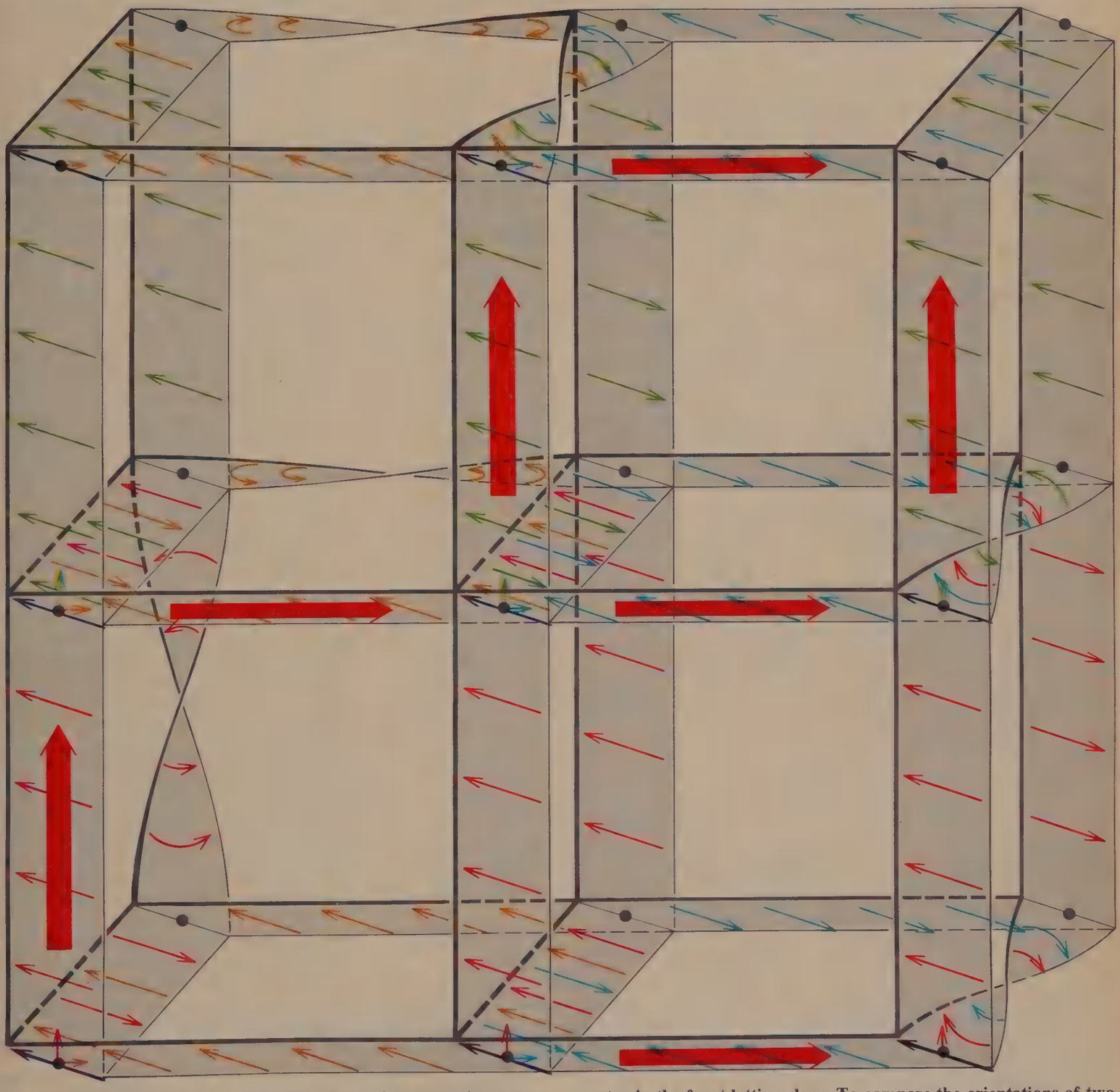
The original impetus for the quark model was the need to bring order to the large number of particles that exhibit strong interactions, or in other words those subject to the strong force. The proton and the neutron are members of this class, and indeed it is the strong force that binds them in an atomic nucleus. The existence of many other strongly interacting particles has been inferred from the decay products of collisions in accelerators. Most such particles live for an extremely short time, as short as 10^{-24} second, before they decay into other particles. All particles that are subject to the strong force are called hadrons, from the Greek adjective hadros, meaning robust or heavily built.

In 1962 Murray Gell-Mann of the California Institute of Technology and Yuval Ne'eman of Tel-Aviv University proposed a scheme for classifying the hadrons in symmetrical patterns. The scheme was based on the mathematical theory of groups and was called the eightfold way. A short time afterward Gell-Mann and, independently, George Zweig, also of Cal Tech, proposed a physical interpretation of the eightfold way. The mathematical classification could be explained by assuming that all hadrons are built up of more fundamental constituents, which Gell-Mann called quarks.

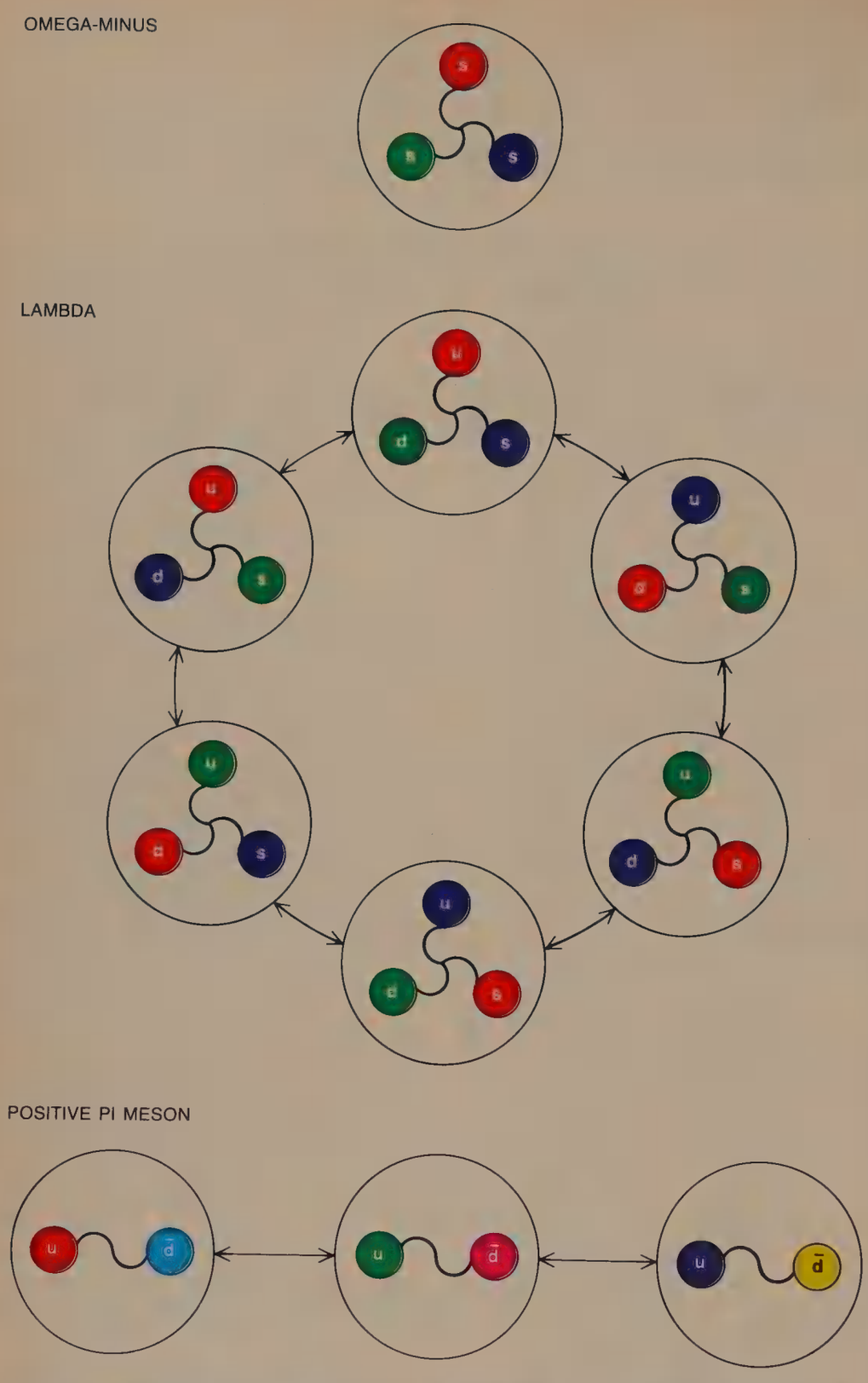
At the time every known hadron could be understood as some combination of three basic quarks (and their corresponding antiquarks): the up or

u quark, the down or *d* quark and the strange or *s* quark. The proton, for example, is a combination of two *u* quarks and a *d* quark, whereas the neutron is a combination of a *u* quark and two *d* quarks. The positively charged pi meson is a combination of a *u* quark and a *d* antiquark. Since the quark hypothesis was put forward more hadrons have been discovered and it has become necessary to add at least two more quarks, the charm or *c* quark and the bottom or *b* quark, to the catalogue of elementary particles. Nevertheless, the quark model remains a highly successful classification scheme: more than 100 hadrons are known, and they can all be described in terms of the quark model.

In spite of the success of the model in classifying hadrons, certain features ascribed to the quarks initially made the



LATTICE OF POINTS is a schematic way of representing space and time; with such a lattice physicists seek to understand the field of force that gives rise to the permanent confinement of quarks. Quarks and antiquarks, which are thought to be the constituents of the neutron and the proton and all other particles subject to the strong nuclear force, have never been observed in isolation. Instead they are bound together in triplets or pairs. The field responsible for the confinement is called the chromoelectric field, and it is classified mathematically as a gauge field. A gauge is analogous to a ruler or a pointer; it is used to compare physical quantities defined at different lattice sites. The length of the ruler and the direction of the pointer can change as they are moved about in space-time. Particles such as quarks lie only at the points, or vertexes, of the lattice. In the illustration the variables that specify the state of a particle are represented by purple arrows, which must point in one of two directions at each vertex in the front lattice plane. To compare the orientations of two purple arrows one of them must be moved next to the other. A set of rules, represented in the illustration by the bands that join neighboring vertexes, must be defined in order to specify the changes made in the orientation of the arrows during transport. The set of rules is a gauge field; it enables an arrow to be transported upward or to the right of a vertex, around a square array of four vertexes called a plaquette and back to its starting point. (The transported arrows are colored green, yellow, blue or pink.) Any set of bands that returns the transported arrow to its original orientation represents the lowest energy state of the gauge field. A set of bands that resembles a Möbius strip, however, cannot be untwisted: a transported arrow returns to its starting vertex with the opposite orientation. The energy of the gauge field defined on the lattice is stored in such plaquettes. The red arrows represent the lines of force to which the gauge field gives rise.



COLOR is a quantum-mechanical property of quarks that was introduced to reconcile the quark model with the exclusion principle of Wolfgang Pauli. The principle states that no two quarks within a small region of space can occupy the same quantum-mechanical state. Before the introduction of the color hypothesis the quark model seemed to predict the existence of particles in which the principle is violated. For example, the constituents of the omega-minus particle are three strange or *s* quarks that were thought to be in the same quantum-mechanical state. The paradox is resolved by assuming that each of the three *s* quarks takes on one of three colors, such as red, purple and green. Since the quarks differ in color, they are in different quantum-mechanical states and the exclusion principle is saved. Color is never observed in isolation, and so the colored quarks must be combined in such a way that the omega-minus particle is colorless (or white). The color hypothesis states further that all possible combinations of the quark colors are equally likely, as long as the composite particle remains colorless. For example, the lambda particle, which is made up of three quarks, can occupy any one of six colorless states with equal probability, namely the six permutations of the three colors red, purple and green. The positive pi meson, which is made up of a quark and an antiquark, can occupy any of three colorless states with equal probability: red and cyan (antired), green and magenta (antigreen) or purple and yellow (antipurple). The confinement of quarks and of color is represented here by showing the quarks linked to one another by an unbreakable string. The quarks can move about almost freely inside a particle as long as they stay within the bounds of the string.

physical reality of quarks difficult to accept. The most fundamental problem is the failure to detect a free quark. The proton and the neutron are strongly bound in the atomic nucleus, yet given enough energy in a nuclear collision they can be set free. Any theory that describes the interactions of quarks, however, not only must account for their binding into hadrons but also must lead to permanent confinement.

Almost equally unsettling was the fact that in certain hadrons the quark constituents seemed to violate a fundamental principle of quantum mechanics, namely the exclusion principle of Wolfgang Pauli. The exclusion principle applies to a broad category of particles including the quarks and states that no two such particles within a small region of space can simultaneously occupy the same quantum-mechanical state. In practice the principle implies that two quarks of the same kind, say two *u* quarks, cannot form a hadron unless they have opposite spins. The spin of a quark is like the spin of the earth, except that the quark's spin is quantized: it can assume only one of two values. Hence in any group of three quarks there must be at least two with the same spin.

There are several hadrons in which three identical quarks must approach one another closely enough to bind together. The omega-minus particle is one such hadron. It was predicted by the quark model, and its subsequent discovery in 1964 by Nicholas P. Samios, Ralph P. Shutt and their collaborators at Brookhaven gave strong support to the model. On the other hand, the omega-minus was also quite puzzling because the quark model predicts it must be made up of three *s* quarks, whose close association seemed to violate the exclusion principle. For these reasons and others physicists initially preferred to regard the quark as a mathematical convenience; the question of its physical existence was temporarily set aside.

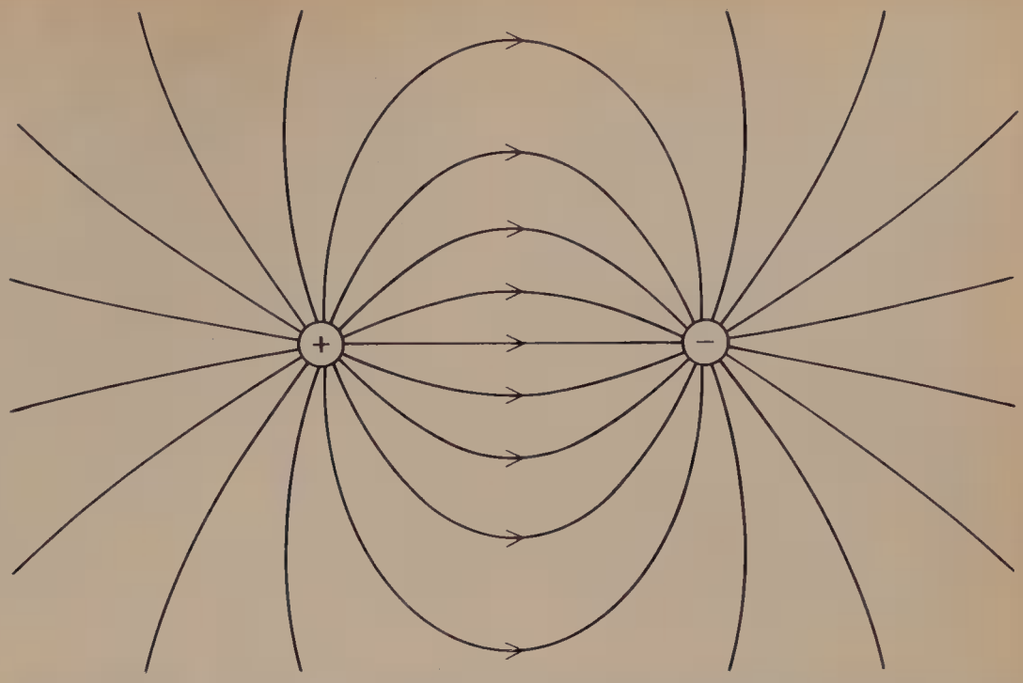
The quark model and the exclusion principle were reconciled as a result of ideas developed by Oscar W. Greenberg of the University of Maryland at College Park and, independently, by Moo-Young Han of Duke University and Yoichiro Nambu of the University of Chicago. What is needed is to assume that each kind of quark can exist in any of three states. For example, if an *s* quark in state *A* is combined with an *s* quark in state *B* and an *s* quark in state *C* to form the omega-minus particle, the exclusion principle is saved. In order to label the states of a quark physicists have whimsically taken to calling them by the names of colors: a quark can come in the three colors red, purple and green, and an antiquark can come in the three complementary colors cyan

(antired), yellow (antipurple) and magenta (antigreen). The prefix "chromo" in quantum chromodynamics refers to the color terminology.

The introduction of color had to be supplemented by another hypothesis if the successful classification of hadrons was to be maintained. Although the new color degree of freedom made possible a quark model of particles such as the omega-minus, it also led to a multiplication of other hadrons. The lambda particle, for example, is made up of a *u*, a *d* and an *s* quark; if each quark can exist in any of three colors, it would seem there should be nine lambda particles, one for each color combination, rather than the single particle that is observed. To avoid such redundancy one adds the hypothesis that the quarks in a hadron can assume only those combinations of colors that leave the hadron colorless, or white, if the rules of color addition (with ordinary light) are assumed. The three quarks in a proton or a lambda particle must include one red, one purple and one green, whereas the quark and the antiquark in a pi meson can be red and cyan, purple and yellow or green and magenta. Because the "total" color is always the same, in the quantum-mechanical sense that each colorless state can occur with equal probability, there is effectively only one lambda particle and only one positive pi meson.

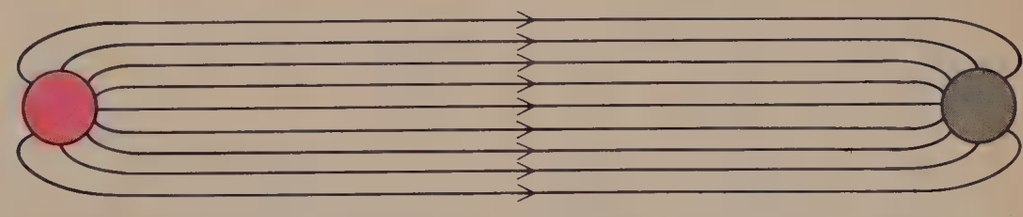
In the late 1960's strong evidence that the quarks in hadrons are real particles instead of mere mathematical entities came from a variety of experimental results. Of notable importance was a series of experiments done at the Stanford Linear Accelerator Center (SLAC) by Jerome I. Friedman and Henry W. Kendall of the Massachusetts Institute of Technology and Richard E. Taylor of SLAC. High-energy electrons were directed against a fixed target of protons in order to probe the protons for internal structure. By examining the decay products of the collisions it was possible to show that inside the proton there are constituents with all the properties attributed to quarks.

Moreover, although no free quarks were detected, the experiments showed that within the proton the quarks are in a nearly free state of motion. This result was quite puzzling: how could forces strong enough to keep quarks permanently bound together also allow them to move about almost freely when they are at close quarters inside a proton? The three quarklike objects in the proton appeared bound to one another like the three stones tied together in a bola, the South American hunting device. The stones in the bola move freely as long as they remain within the limits of the connecting string; the string, however, keeps them from flying apart.



LINES OF FORCE of the electromagnetic field spread out in space. The field shown is the one generated by two particles of opposite electric charge; although the lines of force are densest in the region between the particles, they extend in other directions as well. The intensity of the field at any point (that is, the strength of the force "felt" at a given point by a unit electric charge) is proportional to the number of lines crossing a surface of unit area, orthogonal to the lines of force, that passes through the point. The electromagnetic force that is generated by a single source of electric charge diminishes as the square of the distance from the source.

The new experimental evidence for quarks combined with the introduction of color gave strong impetus to the formulation of a theory of quark dynamics. Color could serve as a source for a new field called the chromoelectric field, which would give rise to a new kind of interaction among colored particles. In 1973 H. David Politzer of Cal Tech and, independently, David Gross of Princeton University and Frank Wilczek of the University of California at Santa Barbara realized that a dynamic interaction based on the chromoelectric field would lead to a progressively weaker force between quarks as they approached one another. The prediction explained the almost free motion of quarks inside protons that had been observed in the SLAC experiments. It was then conjectured that the same interaction could be responsible for the confinement of quarks, although at the time there were no theoretical results to support such an appealing idea. Nevertheless, once the chromoelectric field is introduced the rather ad hoc assumption that all hadrons are colorless and the observed confinement of quarks can be understood as two aspects of the same phenomenon. If one quark, say a red one, is pulled away from a hadron, both the quark and the fragment left behind are colored. If color, like electric charge, is the source of a field, there could be an attractive force between the two colored fragments. Confinement of the quarks might then result if the attraction between the two



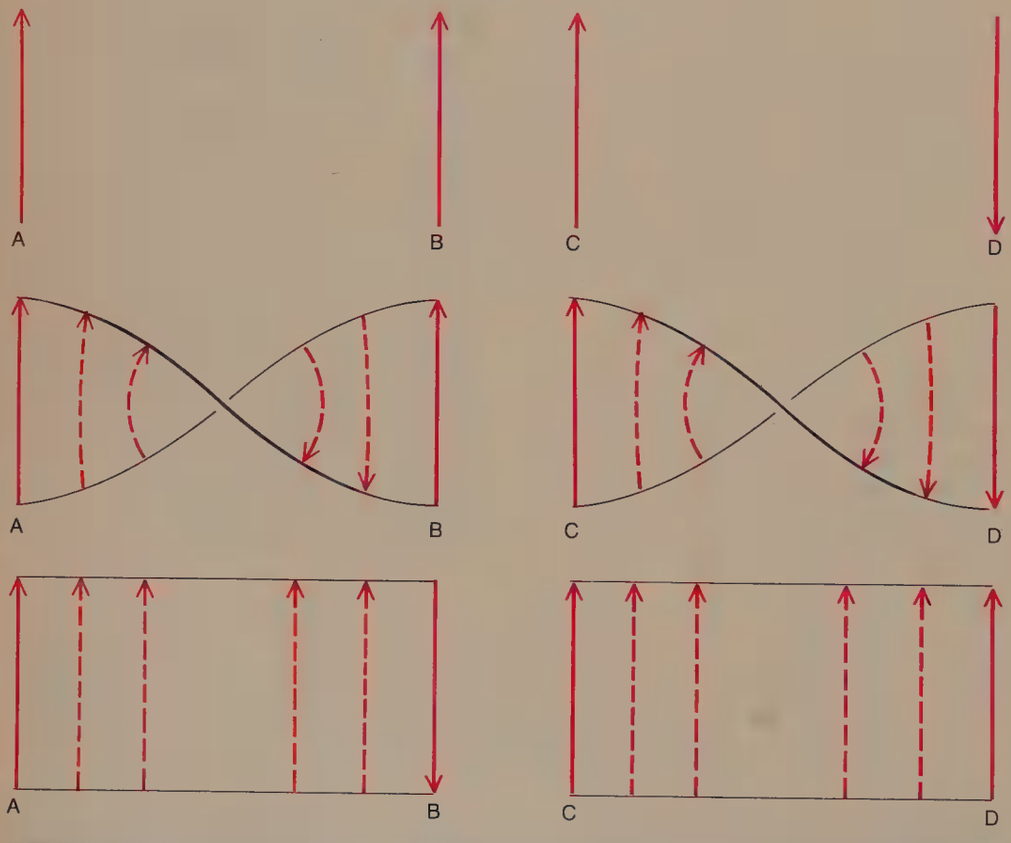
COMPRESSION of the lines of force between two particles into a thin tube of uniform cross section would make the force between the particles constant, regardless of the distance between them. A surface of unit area orthogonal to the tube would always meet the same number of lines of force, no matter where the surface was placed along the tube. Because the force that binds the particles remains constant, increasing the separation of the particles by a given increment would always require the same amount of energy, no matter how far apart they were at the outset. An infinite amount of energy would be needed to free one of the particles from the other. Such compression of the lines of force would therefore lead to the permanent confinement of the two particles. If the radius of the tube is negligible, the bundle of lines of force resembles a string. A model explaining the confinement of quarks by this principle was formulated mathematically in 1968 by Gabriele Veneziano of the European Organization for Nuclear Research (CERN) and interpreted as a string by Yoichiro Nambu of the University of Chicago. The idea that the string could be physically realized as a bundle of lines of force was proposed in 1973 by Holger B. Nielsen and Paul Olesen of the Niels Bohr Institute in Copenhagen.

fragments is so strong that it is impossible to separate them beyond some limit.

In the early 1970's a dynamic model independent of the quark model was developed to account for certain properties of hadrons that had not been explained by quarks. According to the dynamic model, the hadron is not a point or a spherical particle but can better be understood as a string. The string can rotate or vibrate in ways prescribed by the laws of relativistic dynamics, and its end points are required to move at the speed of light. Calculations showed that the force acting along the string must be enormous: about 14 tons. The quantized vibrations of the taut string give rise to various states that could be identified with certain hadrons.

It is evident that the string model of the hadron and the bola analogy can be combined. If three quarks or a quark and an antiquark are placed at the end points of the relativistic string, the tension of the string could explain the permanent binding of the particles. The string model, however, like the original concept of the quark, is itself a mere mathematical abstraction: the string is a one-dimensional object. Could it nonetheless be an approximation of some other structure that is more acceptable from a physical point of view?

In 1973 Holger B. Nielsen and Paul Olesen of the Niels Bohr Institute in Copenhagen pointed out that the string could be interpreted as a bundle of lines of force of a suitable field. In an electromagnetic field, lines of force have the familiar patterns made by iron filings on a sheet of paper held over a magnet. The intensity of the field is proportional to the density of the lines of force. Thus when the lines of force spread out, as they do at points increasingly distant from the poles of an ordinary magnet, the intensity of the field diminishes. If the lines of force are squeezed into a tube of uniform cross section, however, the intensity of the field remains constant all along the tube. The force needed to separate a quark and an antiquark at opposite ends of such a tube would also remain constant no matter how far apart the two particles were placed. In order to liberate one of the quarks an infinite amount of energy would have to be supplied.



DIRECTIONS OF TWO ARROWS defined only at the vertexes of a lattice cannot be compared unless there is some way to transport the arrows from one vertex to the other. In the diagram at the top the arrows at vertexes *A* and *B* seem to point in the same direction, whereas the arrows at *C* and *D* seem to point in opposite directions. The comparison, however, presupposes the existence of the paper on which the diagram is drawn, which acts as an intermediary along which one arrow can be moved next to the other. One might just as well assume that the points in each pair are connected by a half-twisted ribbon of paper, as in the middle diagram; the comparison then gives the contrary result. Without a gauge field to supply a rule of transport from point to point the orientations of the arrows at adjacent points cannot be compared. The diagram at the bottom shows that the result of a comparison need not change when the direction of an arrow is reversed at any point; the comparison remains valid if the ribbon is correspondingly twisted or untwisted. The property of the gauge field that makes it possible to compensate for a change in the direction of any individual arrow is called local gauge invariance.

The quantum-mechanical reality of quarks and strings requires that the lines of force associated with the color interaction of quarks act quite differently from the lines of force associated with the electromagnetic interaction of electrically charged particles. Because both forces propagate in the vacuum, one might assume that any differences between them would be caused by the intrinsic nature of the forces themselves and not by the interaction of the forces with the vacuum. In classical, or Newtonian, mechanics the assumption would be sound; indeed, there can be no interaction between a field and the classical vacuum because the classical vacuum is by definition a state with no matter and no energy in it. In quantum mechanics, however, even the vacuum has a structure, which can alter the propagation of fields and forces.

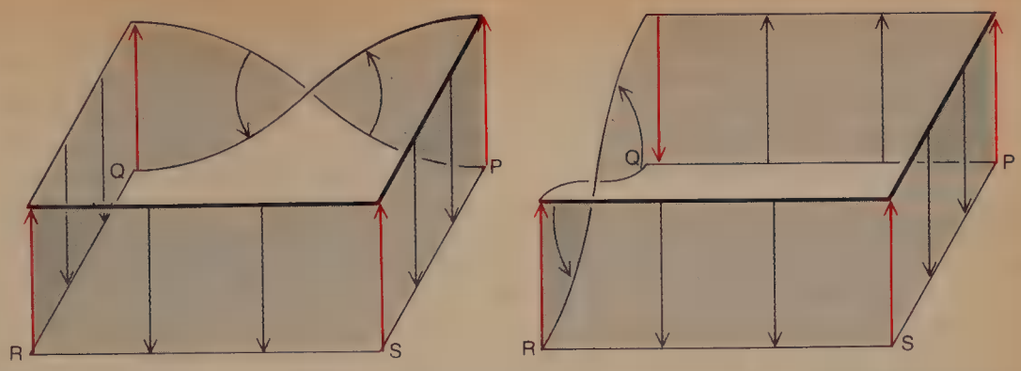
The structure of the vacuum is a consequence of the uncertainty principle of Werner Heisenberg. One version of the uncertainty principle states that for any physical event there is an uncertainty about the energy released during the event that is related to an uncertainty about the exact time of its occurrence. More precisely, the product of the uncertainty about the energy and the uncertainty about the time is not less than some numerical constant. For an event confined to an extremely short interval there is a correspondingly large uncertainty about its energy. During any short interval, therefore, there is a substantial probability that the quantum-mechanical vacuum has some nonzero energy.

The energy of the vacuum can manifest itself in the spontaneous creation or annihilation of a particle and its antiparticle or in the appearance and disappearance of an electric or a chromoelectric field throughout various regions of space. Such variations of a quantum field are called fluctuations. In the electromagnetic field between two electrically charged particles, for example, the presence of quantum fluctuations implies that the interactions between the two charges are not strictly determined by the classical field predicted by Maxwell's equations. Instead the measured electromagnetic field is the average of all the fields that can be generated by the quantum fluctuations, weighted according to the probability that a given fluctuation will occur.

For most practical applications of electrodynamics the effects of the quantum fluctuations are very small. A measurement of the field between two macroscopic electrically charged objects would be consistent with the value predicted by the classical theory. In high-energy collisions of charged particles, however, the quantum-mechanical fluctuations become much more important

and must be taken into account to calculate the electromagnetic effects. In the standard method (which for many problems is quite successful) one first calculates the properties of the field in the classical vacuum. One then builds on the result of the classical calculation by correcting it for quantum-mechanical fluctuations of a progressively higher complexity, according to what is called a perturbative expansion. In the quantum theory of electromagnetism the larger, or the more complex, a fluctuation, the less the probability that it will take place. Hence almost all the corrections to the classical electromagnetic field that must be made in the quantum-mechanical calculations are the outcome of small fluctuations.

One might suppose the properties of the chromoelectric field between two quarks could be deduced in a closely analogous way. It would seem, at least in principle, that a perturbative expansion could give the strength of the field at any point to any degree of accuracy needed. It turns out, however, that the method of perturbative expansion works only if the field calculated for the classical vacuum is the dominant effect. In other words, the method works only if the corrections that must be made to take account of the fluctuations are small and become smaller still as fluctuations of increasing size are considered. For quantum-mechanical phenomena that depend primarily on the effects of large fluctuations the perturbative expansion does not converge, that is, the series of calculations does not approach a constant, finite value. Such phenomena are said to be nonperturbative. The confinement of the lines of force of the chromoelectric field between two quarks, and hence the permanent binding of the two quarks, is a nonperturbative phenomenon.



FRUSTRATED PLAQUETTE is given a unit twist by the four gauge fields that make up its sides. The gauge fields, represented by ribbons, can be twisted or untwisted locally to compensate for the reversal of an arrow at any point, but the overall twist in the four ribbons of the plaquette cannot be removed. The twist between *P* and *Q* (*left*) can be removed by reversing the direction of the arrow at *Q*. The effect, however, is merely to transfer the twist to the part of the ribbon between *Q* and *R* (*right*). No matter which arrows are reversed, or in other words no matter how many local gauge transformations are performed, an odd number of the four ribbons must stay twisted. Hence the plaquette remains frustrated in spite of local gauge invariance.

How, then, can confinement be demonstrated? Workers in theoretical physics recognized that a new approach had to be devised, in which large fluctuations of the quantum-mechanical field are considered at the outset of the calculations. The lattice method suggested by Wilson is such an approach.

The lattice is generally a cubic one and can be thought of as the edges and vertexes of a collection of densely stacked cubes. The lattice extends in time as well as in space, that is, each point on the lattice designates both a position in space and a moment in time. To visualize the lattice one can think of an array of cubes in which two of the axes are labeled with spatial coordinates and the third axis is labeled with temporal coordinates; the full lattice has three spatial dimensions as well as the time axis, and so it is a four-dimensional structure. Between any two neighboring vertexes of the lattice there is a link, which can be pictured as a line connecting the two vertexes. A small square bounded by four links is called a plaquette. In Wilson's formulation the vertexes, links and plaquettes of the lattice are all that is left of ordinary physical space and time.

Links and plaquettes on the lattice must be regarded as entities at a different level of abstraction from the vertexes, or lattice points. Although links and plaquettes are defined by lattice points, there are no additional lattice points along a link or within a plaquette. In other words, space and time on the lattice are quite unlike ordinary space and time, which always include an infinite number of points between any two given points.

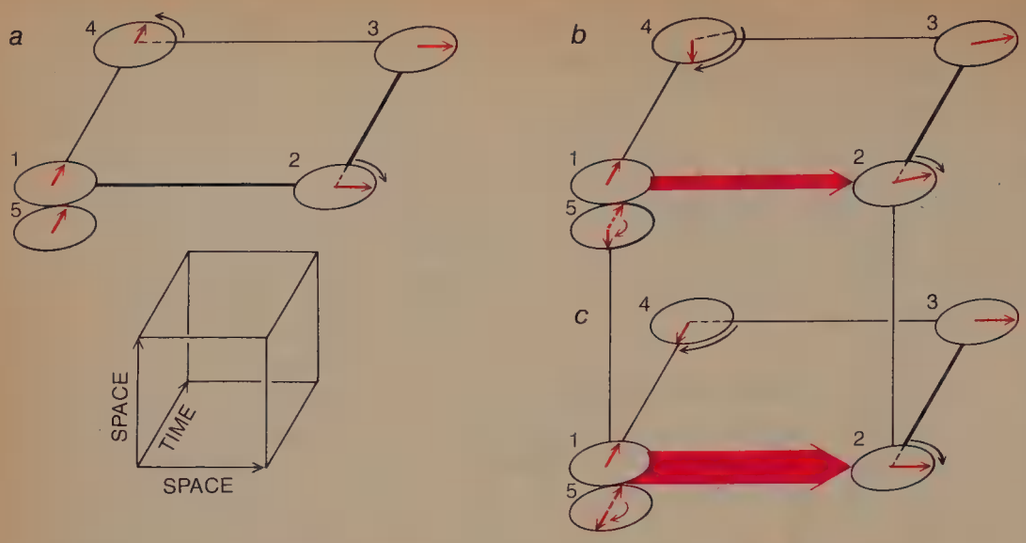
Wilson's introduction of a space-time lattice is not meant to imply that physical processes really take place on a lattice. Space-time, according to all current evidence, is continuous. Instead the lattice represents what theoretical physicists call a regularization, a temporary artifact for making calculations that would otherwise be impossible. Applied to the problem of confinement, the strategy is as follows. All particles are defined only at the vertexes of the lattice, and the strength of the field is defined only along the links of the lattice. (Actually what is defined at each vertex is the probability that a particle will be found there. The probability of finding a particle between two adjacent vertexes is not defined.) When no particles are present, the symmetry of the fluctuations on the lattice dramatically simplifies the calculation of the average electric or chromoelectric field generated by strong fluctuations. The fields, which are vector quantities and therefore have both magnitude and orientation, are just as likely to point in one direction along a link as they are to point in the opposite one. Hence in the vacuum state, with no particles, the mean value of the electric or the chromoelectric field throughout the lattice is zero.

A similar although slightly more elaborate calculation can be done for large fluctuations of the field when a single particle and its corresponding antiparticle are defined on the lattice. On the average the fluctuations of the field again cancel except along the links of the lattice that make up the shortest path between the particle and the antiparticle. The results do not depend on the kind of field defined on the lattice; the particles can be a quark and its antiquark or an electron and a positron. Thus confinement is a natural outcome of defining the field on the lattice.

The next step in the strategy is to remove the lattice and regain ordinary, continuous space-time. The lattice spacing is made progressively smaller so that the vertexes of the lattice become closer and denser in space-time. If the reduction in the lattice spacing proceeds to the mathematical limit, continuous space-time is recovered. At the limit the procedure gives the average field after all the quantum-mechanical fluctuations have been taken into account.

What is gained by introducing the lattice? The strong fluctuations that must be responsible for squeezing together the lines of the color force are considered from the outset along each lattice link. On the other hand, there is a price that must be paid: as long as the lattice is relatively coarse the quantum fluctuations give rise to the confinement of the electromagnetic field as well as the chromoelectric field.

It is well known that electric charges, unlike color charges, do exist in isolation. Since the lattice method predicts the confinement of electric charge, one must be skeptical about the lattice approach unless it can be shown that as the mesh of the lattice is made finer the elec-



GAUGE-FIELD CONCEPT can be generalized by allowing the quantities defined at each lattice point to vary over a continuous range. For example, the direction of an arrow at each point could be allowed to form an arbitrary angle with the vertical. A gauge field makes it possible to compare the angles at different lattice points. If an arrow does not return to its original orientation after a complete circuit of a plaquette, the angular difference between the initial and the final orientations (a quantity called the phase angle) measures the degree of frustration of the plaquette. The electric field is such a gauge field, and the strength of the field in a given direction at any point is measured by the degree of frustration of an associated plaquette. The illustration shows how the frustration of plaquettes is associated with a single component of the electric field, namely the spatial component in the left-right direction. Strictly speaking, the lattice extends in four dimensions, the three spatial ones and time, but one can visualize the lattice as having only two spatial coordinates and a temporal one. As a dial is transported counterclockwise around a plaquette that includes the left-right spatial component and the time dimension, the gauge field causes an arrow on the dial to rotate. If the arrow returns to its starting position after the dial makes a full circuit of the plaquette, the strength of the field is zero in the spatial direction and there is no field line (*a*). As the phase angle increases, the strength of the field along the spatial direction increases as well (*b–c*), which is represented by the increasing thickness of the field lines (*colored arrows*) along the spatial coordinates. Mathematically the rotation of the arrow on the dial can be identified with the rotation of an arrow in the plane of complex numbers, that is, numbers that have both a real part and an imaginary part.

tromagnetic field begins to act in accordance with its well-established properties. In other words, what one would like to show is that at some stage in the shrinking of the lattice the electromagnetic lines of force break out of their confinement to a line of lattice links, whereas the chromoelectric lines of force remain squeezed together all the way to the limit of continuous space-time. It is not a straightforward matter to demonstrate that things happen in precisely this way, but in the past few years the demonstration has been achieved by making elaborate numerical calculations with the aid of high-speed computers.

Theoretical physicists express the fundamental difference between the electromagnetic field and the chromoelectric field by saying that QED is an Abelian gauge theory whereas QCD is a non-Abelian gauge theory. The terms refer to the Norwegian mathematician Niels Henrik Abel. The distinction between Abelian and non-Abelian is drawn from the mathematical theory of groups, which describes the symmetries inherent in a sequence of operations, such as a sequence of rotations. If the operations that are members of a group can be carried out in any sequence with the same final result, the group is Abelian. For example, the group of rotations about a single axis is Abelian, because such operations have the same effect regardless of their sequence. On the other hand, if the sequence in which two or more operations are carried out does affect the final outcome, the group of operations is a non-Abelian one. The rotations of a cube about its three axes form a non-Abelian group: when the cube is turned about a vertical axis and a horizontal one, the result depends on which operation is done first.

In order to understand how ideas from group theory apply to QCD and QED, one must understand the concept of a gauge field. The concept can best be illustrated for isolated points of space and time arranged in a lattice. Particles can rest on any of the lattice vertexes or hop from one vertex to another; as the particles move through the space-time lattice they can change their state, where a state is defined by quantities that can vary over a certain range of values. I shall make the simplifying assumption that the state of a particle is described by just one variable, which can take on exactly two values. For example, at each vertex of the lattice there might be a variable whose value is either $+1$ or -1, indicating the sign of the electric charge. The two possible values can be represented by an arrow that points either up or down.

In describing how interactions propagate in space and time it is essential to be able to compare the values of variables at neighboring points. To compare the length of two objects in separate regions of space one needs a measuring stick, or gauge, that can be moved next to one object, marked and then moved to the second object, where the comparison is made. Similarly, on the lattice one must be able to compare the orientation of the arrows at neighboring vertexes. At first such a comparison seems trivial. Suppose two vertexes of the lattice are represented on a sheet of paper and at each vertex there is an arrow directed upward. Is it not obvious that the two arrows point in the same direction? The question, however, presupposes the existence of the sheet of paper on which the lattice is drawn. The paper acts as an intermediary that allows the two orientations to be compared. In a sense, one transports the arrow from one vertex to the other with the eye, and one concludes that the two arrows had the same orientation before the transport because after the transport they match.

Suppose the sheet of paper between the two arrows had been given a half twist [*see illustration on page 58*]. An upward-pointing arrow transported along the twisted paper would point downward when it reached the second vertex. Because the direction of an arrow is defined only at the individual, isolated vertexes of the lattice, there is no way to decide which of the two methods of transport is correct. Indeed, without the sheet of paper or some similar assumption about the effects of transport no comparison of directions at different vertexes is possible.

In field theory a gauge is any standard of measurement, analogous to the distance between two marks on a metal bar or the direction of an arrow with respect to a dial, that can change under the influence of a field as the gauge is moved about in space and time. A field that can effect such changes is called a gauge field, and it specifies explicitly the assumptions that must be made about the transport of the gauge. In my simple example the gauge field is a set of rules for transporting the arrows along the links of the lattice from one vertex to the next. One can think of the gauge field as a ribbon that connects neighboring lattice sites and thereby allows the arrows at different points to be transported and compared.

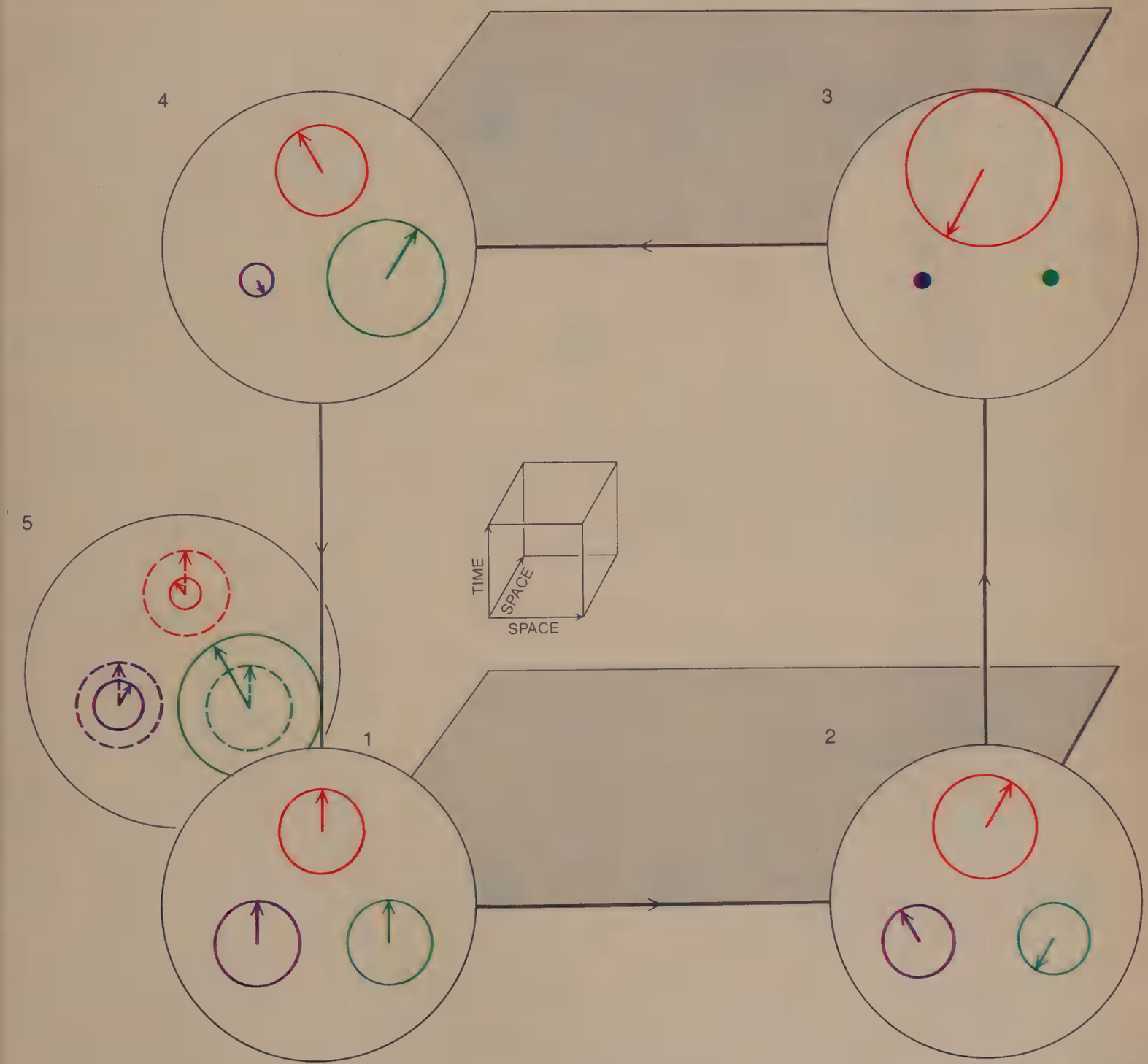
It is important to note that the arrow representing the state at any lattice vertex can be reversed, provided the ribbon representing the gauge field is correspondingly twisted or untwisted. If the arrows and the ribbons are adjusted in

coordination, the result of any comparison between state variables does not change, nor does the physical information represented by the system. The possibility of reversing the arrows without modifying the physical information is called local gauge invariance.

Because of local gauge invariance one might think a gauge field is nothing but an unnecessary complication. What is the point of introducing the twisted ribbon if it can then be untwisted by a local gauge transformation? The objection is valid as long as one considers only pairs of lattice points. The importance of the local gauge invariance becomes apparent, however, when one considers a plaquette on the lattice: a set of four points in a square, connected along the lattice links by four ribbons that constitute the gauge field. Suppose one of the ribbons is twisted and the other three are not [*see illustration on page 59*]. The single twisted ribbon can be untwisted (and one of the arrows can be reversed), but not without introducing a half twist in one of the other three ribbons. No matter which arrows are reversed and no matter how many local gauge transformations are carried out, an odd number of the four ribbons must be twisted. A plaquette whose ribbons cannot all be untwisted is called a frustrated plaquette.

The frustration of the plaquette manifests itself in another way. An upward-pointing arrow that is transported along the ribbons around the frustrated plaquette will return to its starting vertex pointing down. Thus on the lattice the frustrated plaquettes are those that generate a mismatch in orientation when an arrow is transported all the way around.

It is not difficult to apply the idea of a

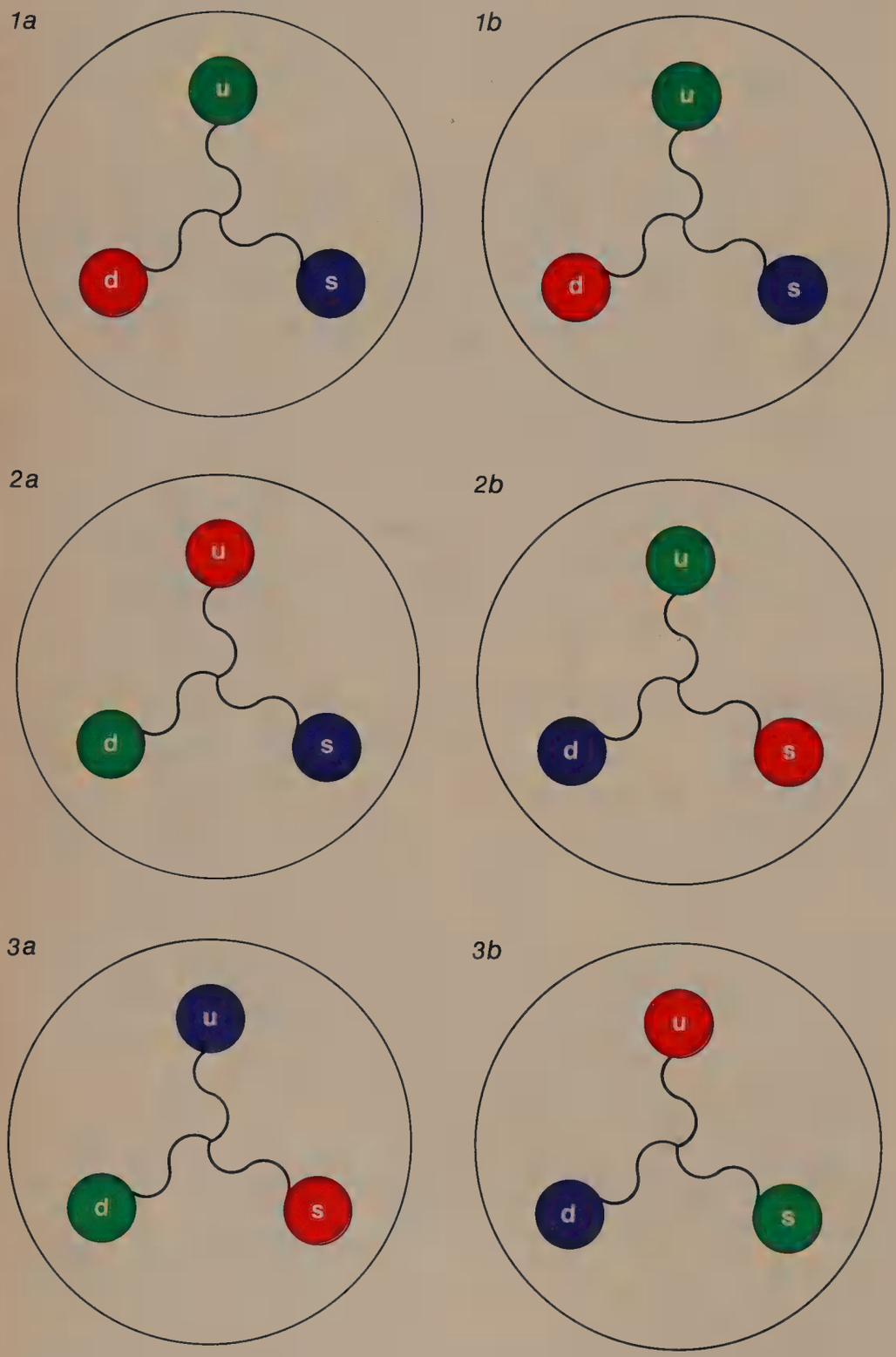


CHROMOELECTRIC FIELD is a gauge field similar in principle to the electromagnetic field but more complicated mathematically. At every point on a lattice there are three arrows instead of one; they correspond to the color charges of a quark. Moreover, the color gauge field affects not only the direction of each arrow but also its length. The lengths of the arrows are not independent of one another: the square root of the sum of the squares of the lengths must equal 1. The strength of the chromoelectric field along each link of the lattice is dependent on the phase angles and the change in the configuration of the three arrows after a complete circuit of a plaquette.

frustrated plaquette to ordinary space-time. In physical space-time the ribbons along which the arrows move cannot be visualized directly; just as the sheets of paper that define the gauge field on the lattice are not themselves part of the lattice, so a gauge field defined in ordinary space-time is not itself part of space-time. Mathematically the higher-dimensional abstract space that specifies the rotations of the arrow is called a connection in a fiber bundle [see "Fiber Bundles and Quantum Theory," by Herbert J. Bernstein and Anthony V. Phillips; SCIENTIFIC AMERICAN, July, 1981]. Nevertheless, it is possible without complete mathematical understanding to imagine that an arrow moved about along some closed path in space-time could return to its starting point with its direction changed.

The idea of a frustrated plaquette has an important physical interpretation. In any gauge field the energy of the field resides precisely in the plaquettes that are frustrated. The plaquettes that are not frustrated, in which all the ribbons can be untwisted and all the arrows can be oriented in the same direction, are associated with the vacuum state of a physical system, the configuration with no energy. Frustrated plaquettes are the sites of the fluctuations in a quantum-mechanical field.

The idea of a gauge field can readily be generalized to more complex situations. The arrows, for example, can be allowed to form an arbitrary angle with respect to a fixed direction rather than being constrained to point up or down. The gauge field now specifies the angle by which the arrow rotates when it is transported along a link. Consequently the frustration of a plaquette can take on a continuous range of values, measured by the angular difference in the direction of the arrow after it is transported around the entire plaquette. The degree of frustration of a plaquette, expressed in suitable mathematical units, is called its action. The action of the entire system is the sum of the actions of all the individual plaquettes.



GROUP OF COLOR TRANSFORMATIONS in the chromoelectric field is classified mathematically as a non-Abelian one. A non-Abelian group is an abstract set of operations (together with the objects on which the operations are carried out) in which the final outcome of several operations is dependent on the sequence in which they are done. For the chromoelectric gauge theory the operations are interchanges of colors among quarks. In the illustration two interchanges of color are applied to a given quark configuration (*1a, 1b*). At the left an exchange of red and green (*2a*) is followed by an exchange of purple and red (*3a*). At the right the order of the two exchanges is reversed (*2b, 3b*). The final configurations of the quarks are not the same.

It turns out the electric field is a gauge field that specifies the continuous rotations of an arrow about a single axis, although this fact is not apparent in any but the most unified and sophisticated formulations of the concept of the electric field. It is more common to think of the electric field as a set of vectors, with one vector at each point in space giving the magnitude and direction of the field at that point. Understood according to the more comprehensive framework of the gauge theory, however, the magnitude of each vector is directly proportional to the frustration of an associated plaquette.

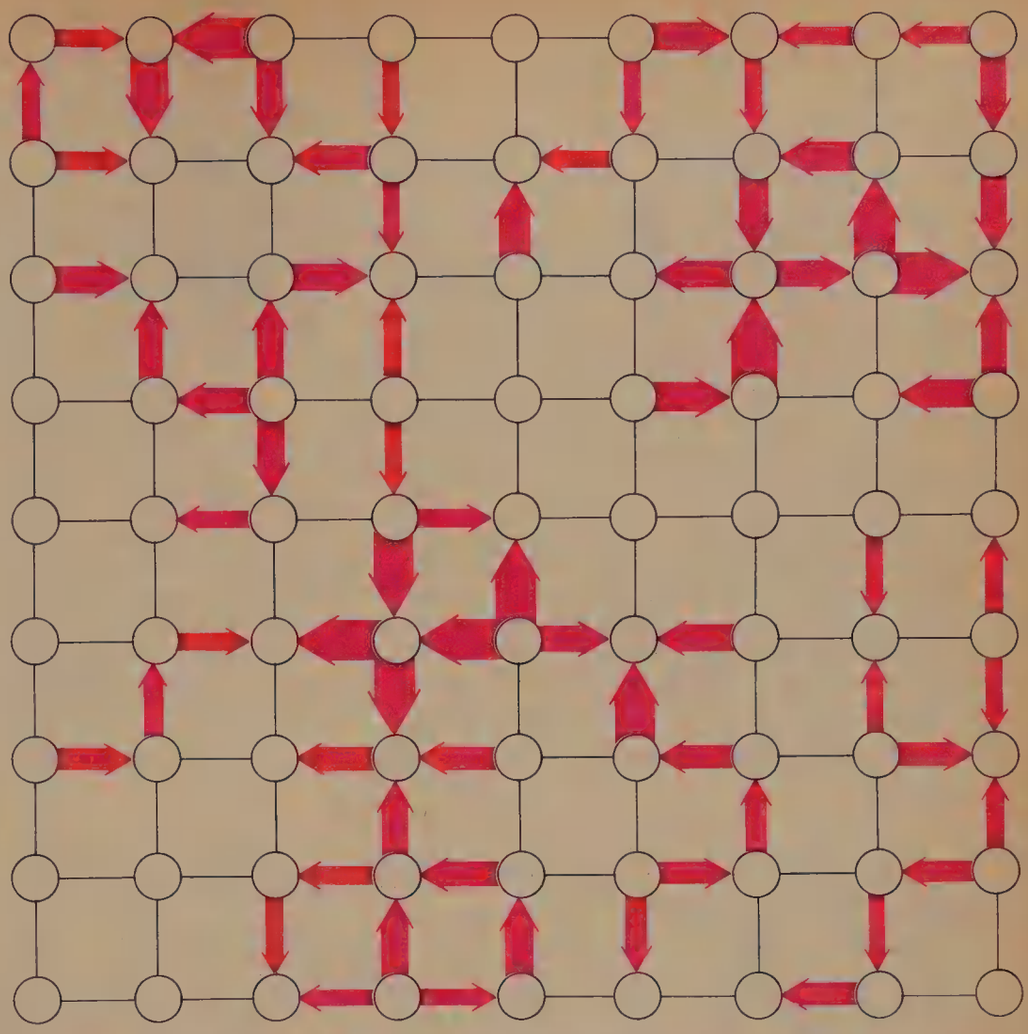
The rotating arrow of the gauge field associated with electromagnetism actually represents a complex number (one with both a real part and an imaginary part), which changes in value as the arrow is carried around the plaquette. (The transported arrow and the complex number must not be confused with the vector that represents the electric field itself at each link on the lattice.) The plaquette is a loop in space-time, and so it is necessary to imagine the transported arrow as moving forward and backward in time as well as in space. The arrow is carried along a certain axis in space, say the positive x axis, then forward in time, then back to its starting point on the x axis and finally backward to its starting point in time. The amount

by which the direction of the transported arrow changes as a result of the transport is called the phase angle. According to gauge theory, the magnitude of the x component of the vector that ordinarily gives the strength of the electric field at a point is a measure of the phase angle generated around a space-time loop that begins in the x direction. The quantum-mechanical fluctuations at a point in the electric field can therefore be thought of as fluctuations in the amount of rotation an arrow would undergo if the arrow were transported around a plaquette extended along one spatial dimension and along the temporal dimension [*see illustration on page 60*].

It is difficult to suppress a feeling of unreality when one is asked to regard the electric field—an entity that can become quite tangible when one gets a shock—as the abstract space of phase rotations. Nevertheless, as one goes more deeply into the study of the physical world, tangible facts and mathematical concepts become intertwined. The abstract idea of phase space at last makes contact with the distinction between QED and QCD, that is, with the distinction between an Abelian gauge theory and a non-Abelian one. The fundamental group operations in QED are the rotations of phase angles. The result of two successive phase rotations, say one of 30 degrees and one of 50 degrees, does not depend on the sequence in which they are done. The net result in either case is a rotation by 80 degrees. Such an outcome is characteristic of an Abelian group and QED is therefore an Abelian gauge theory.

If the state variables defined at the vertexes of a lattice are the colors of quarks, the transported gauge is made up of three arrows instead of just one, and the arrows can vary in length as well as in angle [*see illustration on page 61*]. Each arrow represents one of three color charges; the color charges together determine the color of the quark. As the quark is transported along a link of the lattice it can change its color from red to green; hence a red quark can exchange colors with a green quark, and then the newly green quark can exchange colors with a purple quark. The final result of the exchanges, however, depends on the sequence of events. Since an outcome that depends on sequence is characteristic of a non-Abelian group, QCD is a non-Abelian gauge theory.

The non-Abelian nature of QCD introduces an extra degree of freedom into the fluctuations of the chromoelectric field. Moreover, the existence of three kinds of color also implies that the plaquettes of the chromoelectric field can be frustrated in many more ways than the plaquettes of the electromagnetic field can. It is likely that the extra degree of freedom and the additional disorder of the non-Abelian theory are responsible for the confinement of quarks within hadrons.

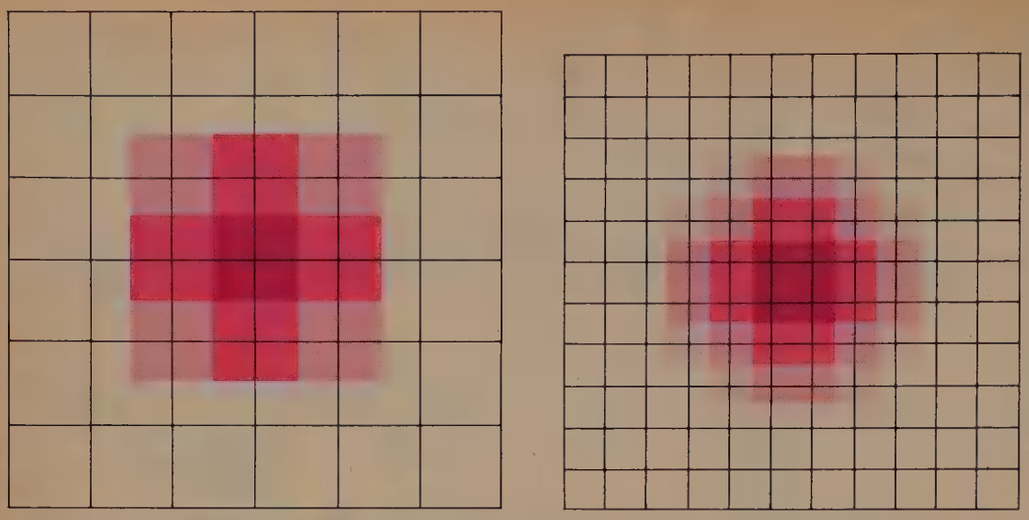


FLUCTUATIONS of the lines of force on a lattice are characteristic of the quantum-mechanical vacuum. The fluctuations are a consequence of the uncertainty principle of Werner Heisenberg. The principle states that the product of the uncertainty about the energy of a system and the uncertainty about the time of an event is not less than a numerical constant. Hence during any short interval there is a correspondingly large uncertainty about the energy of the system, which manifests itself in spontaneous fluctuations of the quantum field. The fluctuations cancel one another in the absence of charged particles, and so a measurement of, say, the electric field in the vacuum would give the value zero because it would determine only the average of all the fluctuations. Nevertheless, the fluctuations of the lines of force in the vacuum must be carefully considered in calculating the field of force between two charged particles.

I have already stated that confinement on a coarse lattice can be demonstrated by a relatively simple calculation for large fluctuations of the quantum-mechanical field. For smaller fluctuations on a finer lattice, however, the calculation of the field becomes much more difficult. The value of any physical quantity measured in an experiment is the quantum-expectation value, which is a weighted average of all possible values the quantity can have. The measured electric or chromoelectric field is the average of all the possible configurations of the field to which fluctuations can give rise. Not all configurations contribute equally to the average, and so each configuration must be weighted, or multiplied by some factor based on the probability of the configuration. In principle, therefore, the quantum-expectation value of a physical quantity defined on the lattice can be calculated in two steps. First the value is calculated for each configuration of the field fluctuations and multiplied by the weight factor for that configuration. Then the products are added and the result is divided by the sum of the weight factors.

Even for a lattice defined over a small volume of space-time, however, the number of possible configurations is so large that a complete summation is out of the question. For the simplest-possible gauge theory, in which the gauge field between two lattice vertexes can be either twisted or untwisted, the number of configurations on a lattice extended for 10 sites in each direction is $2^{40,000}$, or more than $10^{12,041}$.

To calculate the quantum-expectation value of the field on the lattice one must therefore resort to techniques of statistical sampling. The techniques are analogous to the ones employed in opin-



SHRINKING A LATTICE on which particles and fields have been defined must be carried out in order to approximate the effects of the field in ordinary, continuous space and time. As the lattice spacing is reduced, however, the size of a composite particle such as a proton must be preserved, that is, the probability that a quark will be found at a given lattice vertex within the proton must be spread over an increasing number of vertexes. The illustration at the right shows the result of shrinking the lattice at the left to half its previous size. The color density represents the relative probability that a quark will be found at a given vertex. The shrinking can be accomplished by systematically changing the value of the quantity called the coupling constant, which determines the strength of the fluctuations in the quantum-mechanical field.

ion polling. One cannot ask the opinion of every person in the U.S. in order to determine how an issue is perceived by various groups in the population, and so a sample is selected. The result should reflect the actual opinion of the groups if the probability of selecting a respondent from a given group matches the portion, or weight, of the group in the population as a whole.

In a similar way the configurations of the field fluctuations are sampled by a computer program. The computer generates a large number of configurations (but many fewer than $10^{12,041}$), and the probability that a particular configuration is generated is set equal to the quantum-mechanical weight factor for that configuration. The sample of configurations tends to give the same average value for the quantum field as the total population of configurations does.

The importance of a configuration in calculating the quantum-expectation value is determined by its action, which is generally designated S; the weight factor is therefore given by a mathematical function of the action S. An elegant formula for the weight factor in quantum-mechanical systems was introduced by Richard P. Feynman of Cal Tech. First the total action of the configuration is divided by a constant, g^2, then the exponential of this quantity is determined; in other words, the number e (equal to approximately 2.7) is raised to the power S/g^2. The weight factor is inversely proportional to the result. Thus the weight factor is proportional to $1/\exp(S/g^2)$.

The formula for the weight factor implies that the higher the action of a configuration is, the less the configuration is weighted in calculating the average quantum field. Because the action S is not itself exponentiated but instead is divided by g^2, changing the value of g can have a significant effect on the weight factor of a given configuration. If g is large, S/g^2 is small and $1/\exp(S/g^2)$ is larger than it would be if g were small. The quantity g is called the coupling constant; hence when the coupling constant is large, the weight factor of a quantum-mechanical fluctuation with large action is higher than it is when the coupling constant is small.

The idea of a coupling constant may be familiar as a measure of the intrinsic strength of a force. In electromagnetic theory the coupling constant is an important physical quantity with a value of about 1/137. From the rather abstract perspective of the lattice gauge theory, however, g is to be understood as a quantity that takes on a fixed value for a given lattice spacing but can vary with the spacing. Once confinement is demonstrated on a coarse lattice the mesh of the lattice is made finer by carefully decreasing the value of the coupling constant. The process by which the lattice is made progressively finer until continuous space-time is recovered is called renormalization.

Strictly speaking, it is not yet possible to renormalize the lattice for fluctuations of the chromoelectric field, or in other words to shrink the lattice spacing to zero. Nevertheless, it is possible to make the lattice spacing smaller while still keeping it greater than zero and to search for indications that the lines of force do or do not remain confined on finer lattices. As the lattice spacing is reduced all physical objects should remain the same size. On a coarse lattice the probability distribution for a proton might be defined as nonzero across only three lattice spacings. When the lattice spacing is reduced to half its original value, the probability distribution must stretch over six lattice spacings. Since reducing the value of g lowers the probability of lattice configurations that have a large action, the decrease in g has the effect of "zeroing in" for a closer look at the fluctuations of the field, at the lines of force and at the particles defined on the lattice. Thus reducing the value of g makes the proton look larger on the lattice.

The numerical investigation of the properties of a gauge field on the lattice proceeds by limiting the lattice to a large but finite volume. The number of lattice vertexes, links, plaquettes and state variables is therefore finite, although it may be larger than 100,000. Initial values of all the variables are stored in the memory of a large computer. By randomly varying the elements in the starting configuration according to a suitable algorithm the computer generates a sample of as many as 100,000 configurations. Finally it calculates the average quantum-mechanical effects to which the configurations in its sample give rise.

Because of the element of randomness in the calculation, the method is called a Monte Carlo simulation. Before my own work and that of my colleagues on quark confinement the Monte Carlo method had been applied with considerable success to the analysis of the properties of thermodynamic systems, and Wilson had emphasized its suitability for the analysis of lattice gauge theories in quantum mechanics. In 1979, working at Brookhaven, Michael J. Creutz, Laurence A. Jacobs and I first applied Monte Carlo simulation to the study of Abelian gauge theories. We wanted to test whether or not the confinement of particles on the lattice observed at large values of the coupling constant in QED disappears as it should when the continuum limit is approached. The results were spectacular; they showed clearly that at some stage in the reduction of the coupling constant the lines of force on the lattice suddenly undergo a transition. The electric field, which for large values of g is confined to the lattice links between two electric charges, suddenly spreads out all around the charges.

It is useful to compare the sudden deconfinement of the electric field with the sudden change in properties observed when a solid changes to the liquid phase. What Creutz, Jacobs and I observed in our numerical calculations was a phase transition. Everything happened in the computer model as if we had a four-dimensional crystal that we could heat or cool by changing the value of g. At

a certain point the crystal underwent a phase transition; on one side of the transition electric charges are confined, whereas on the other side they are not. Our results were later confirmed independently by an extensive numerical simulation done at the European Organization for Nuclear Research (CERN) by Benny Lautrup of the Niels Bohr Institute and Michael Nauenberg of the University of California at Santa Barbara and by other investigators.

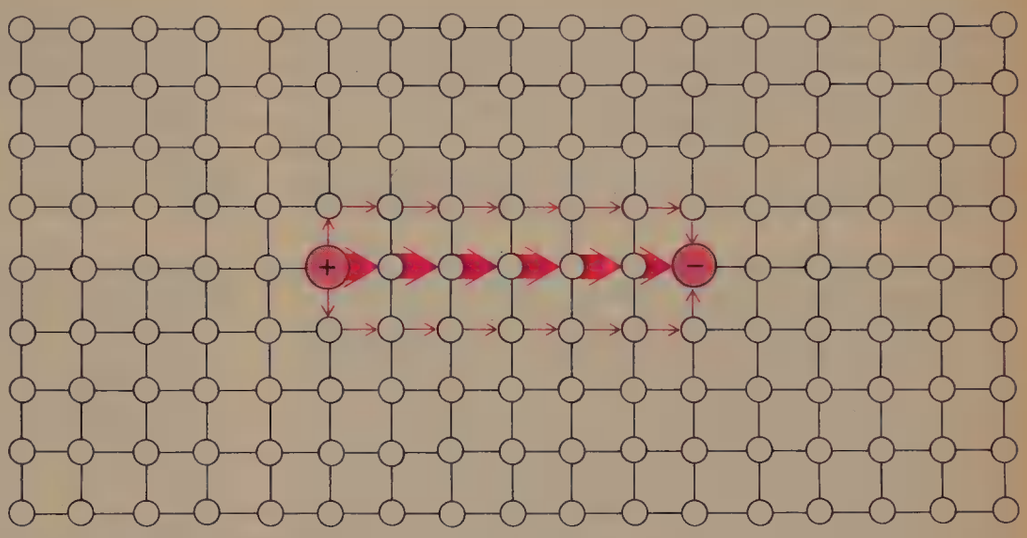
Soon after we got our results for the Abelian lattice gauge theory Creutz extended the Monte Carlo simulation to the non-Abelian model. His results were just as spectacular and physically more interesting; with QED the correct answer was known beforehand and the simulation was a test of the Monte Carlo method rather than of QED, but the outcome for a non-Abelian system was entirely unknown. Creutz's simulation showed that, contrary to the Abelian case, the non-Abelian lattice gauge theory undergoes no phase transition as the value of the coupling constant is gradually lowered. Thus what had long been sought is now finally obtained: a demonstration, albeit by numerical methods, that quantum chromodynamics leads to the confinement of quarks.

Monte Carlo simulation makes it possible to explore the predictions of QCD for many physical processes. For example, after Creutz showed that there is no deconfining phase transition in QCD he was also able to estimate the force holding the quarks together. The result is in excellent agreement with the prediction of the string model, which makes no pretense of describing the dynamics of quarks and hadrons in their full generality. Creutz's value for the string tension was confirmed independently by Wilson, by Gyan Bhanot of CERN, by me and by several others.

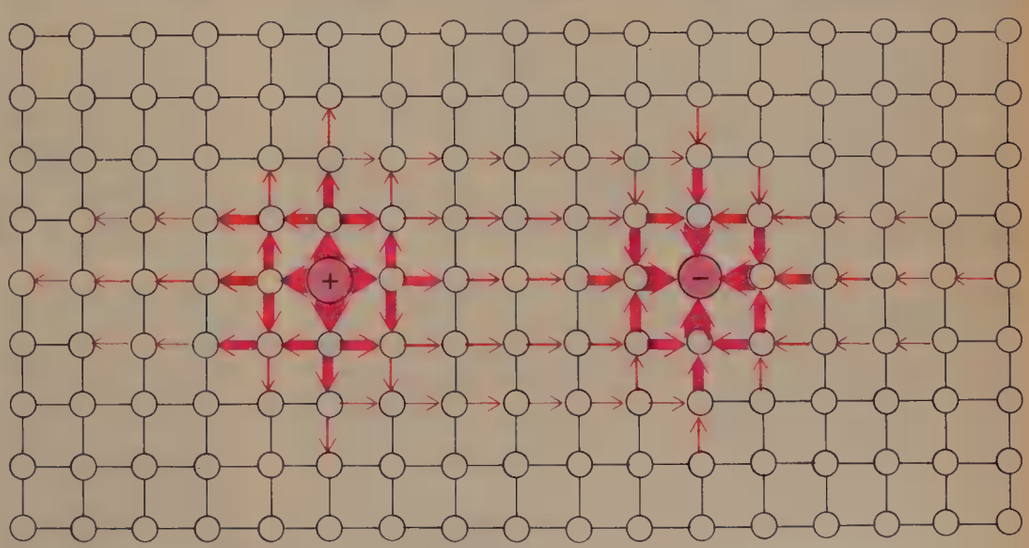
Another prediction of Monte Carlo simulation is the surprising result that at extremely high temperature quarks become deconfined and can move about freely. The temperature must be about two trillion degrees Celsius, much too high to be created in the laboratory but perhaps not too high for free quarks to be found in the interior of hot stars. The prediction also suggests that free quarks existed shortly after the big bang.

The current frontier in the understanding of quark interactions is the numerical simulation of changes in the chromoelectric field as the quarks move about. All the investigations described so far evaluate the chromoelectric field by assuming that the quarks are stationary. It is possible to employ the Monte Carlo method to simulate the quantum mechanics of moving quarks in a gauge field, but the amount of computation required puts the simulation almost out of reach, even for the most powerful computers. Herbert W. Hamber of the Institute for Advanced Study, Enzo Marinari and Georgio Parisi of the University of Rome and I, and independently Donald H. Weingarten of Indiana University, have recently proposed an approximation scheme whereby the computations become feasible. The approximation has been applied to a calculation of the masses of several hadrons, and the theoretical results are in good agreement with experimental values.

Lattice gauge theory has at last brought QCD to the stage where one can calculate its predictions and compare them with experiment. It is an elegant theory, based on relatively simple mathematical concepts yet rich in consequences. For the moment all its verifiable predictions are passing the test of experiment. What remains to be done is to learn how its consequences can be proved by logical deduction. Although a numerical approximation cannot satisfy the need for logical demonstration, it does augur well for such a demonstration in the future. In other branches of theoretical physics there have been many instances in which a definite result, obtained numerically at first, was soon proved by analytic methods. Several of my colleagues and I are confident that QCD will soon pass this test as well.



CONFINEMENT OF A FIELD, whether the field is the electric or the chromoelectric one, can be demonstrated by fairly simple calculations when the field is defined on a coarse lattice and the fluctuations of the lines of force are strong. The strong fluctuations cancel one another everywhere on the lattice except in the region between two charged particles. The lines of force are compressed into a thin tube, and so the two particles cannot be moved indefinitely far apart.



CALCULATION OF THE ELECTRIC FIELD generated by two oppositely charged particles defined on a lattice was done by the author with a Control Data Corporation CDC-7600 computer at the Brookhaven National Laboratory. The illustration shows the outcome of thousands of calculations, done for progressively finer lattice spacings. As the spacing is reduced the lines of force that are confined on the coarse lattice (*see upper illustration on this page*) suddenly undergo a phase transition and break out in all directions. The deconfinement of the electric field on the lattice confirms the plausibility of the lattice approach, because it is well known that the electric field spreads out in ordinary, continuous space. Similar calculations done by Michael J. Creutz of Brookhaven have demonstrated that as the lattice spacing is reduced for the chromoelectric field the lines of force remain confined, just as they are on the coarse lattice. Thus the confinement of quarks can be derived from the theory of the chromoelectric field.

Synthetic Vaccines

A short chain of amino acids assembled in the laboratory to mimic a site on the surface of a viral protein can give rise to antibodies of predetermined specificity that confer immunity against the virus

by Richard A. Lerner

Vaccination is one of man's most significant inventions. It has eliminated smallpox and brought such diseases as diphtheria, poliomyelitis and measles under control. Yet the procedure remains imperfect: the safety of a vaccine cannot be absolutely ensured. Immunization against viral infection is achieved by injecting a virus that elicits antibodies capable of neutralizing the agent of a disease; the injected virus is not itself likely to cause infection because it has been either attenuated or killed. An attenuated virus is one whose virulence in human beings has been reduced by mutation in the course of passage through some different animal host, so that the virus becomes adapted to survival in that animal rather than in human beings. (The cowpox virus, which confers immunity to smallpox, is in effect a naturally attenuated virus.) An attenuated virus is, however, a living organism and therefore a changeable one. It can mutate further, possibly increasing in virulence. Even in the absence of mutation it may possibly have some unknown long-term effect comparable to those of slow-acting "latent" viruses. These possibilities are eliminated if the virus is killed and thereby inactivated, but there have been instances where an incompletely inactivated virus has caused disease.

Other problems are common to both attenuated-virus and killed-virus vaccines. Any facility where viruses are grown to make a vaccine is a reservoir from which an agent of disease can accidentally be disseminated. Moreover, because viruses replicate only in a living system, they are commonly grown in cultured cells or fertilized eggs or are isolated from the blood of infected animals (from infected human beings in the case of the virus causing hepatitis *B*); the cultured cells, eggs or blood may contain undetected substances, notably other viruses, that can contaminate the vaccine. Some vaccines, even some killed-virus ones, must be kept under refrigeration from the time they are prepared until they are inoculated, and maintaining a dependable "cold chain" can be a formidable problem in underdeveloped parts of the world.

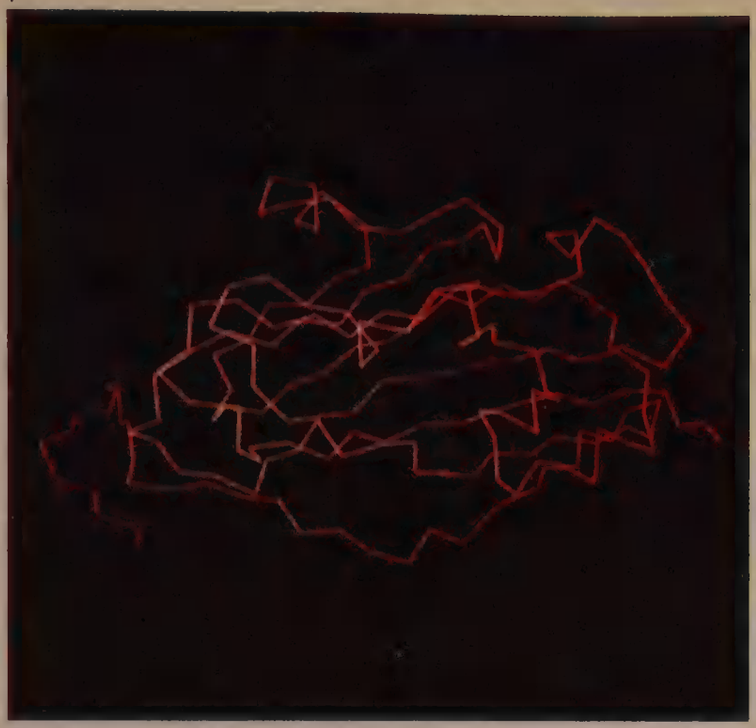
For all these reasons there has been increasing interest in the preparation of synthetic vaccines, which is to say vaccines containing not intact viruses but merely peptides (short protein chains) that have been constructed in the laboratory to mimic a very small region of the virus's outer coat and that can nonetheless give rise to antibodies capable of neutralizing the virus. Experimental vaccines of this kind have now been developed in a number of laboratories, including ours at the Research Institute of Scripps Clinic. In the process much has been learned about the immunological structure of proteins; indeed, the ability to synthesize peptides that elicit antibodies of precise and predetermined specificity may give molecular biologists a battery of tools whose significance may rival the clinical significance of synthetic vaccines.

Before reporting on our experiments and results and on what we have learned about protein structure I should describe the immune system's mode of defense against viral infection. A virus is a small package of genetic information (DNA or RNA) enclosed in a capsid consisting of many copies of a protein or of several different proteins; the capsid may in turn be enveloped in a membrane studded with proteins. Infection begins when a specific site on the surface of the virus binds to a specific receptor on the surface of the cell. The virus penetrates the cell membrane and appropriates the cell's biosynthetic machinery to make many copies of itself, which burst out of the cell and go on to infect more cells. The infective process may be blocked by the immune system if the host has previously been exposed to the same virus or has been immunized against it with a vaccine. In either case certain of the immune-system cells called lymphocytes will have been primed to recognize the virus. When receptors on one of these lymphocytes encounter the virus, an immune response is triggered, one result of which is the proliferation of plasma cells that secrete antibodies against sites on the surface of the virus. The antibodies bind to those sites, coat the surface of the virus and block its attachment to receptors on cells. The virus is neutralized.

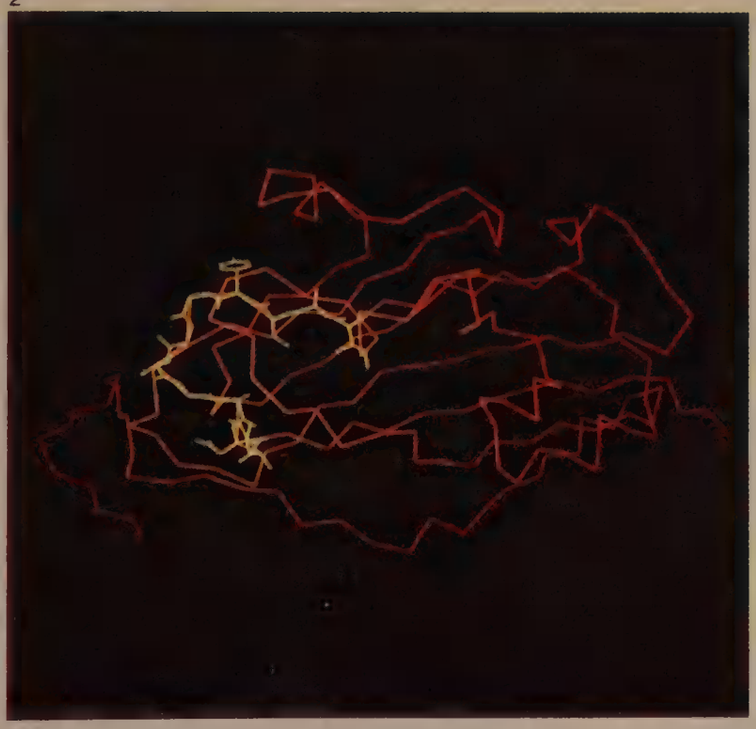
The site that is recognized by a specific lymphocyte and to which the antigen-binding site of a specific antibody subsequently binds is called an antigenic determinant. It is usually a limited patch of a viral surface protein. That being the case, might it not be possible to elicit antibodies capable of neutralizing a virus by injecting as a vaccine peptides that constitute antigenic determinants? The notion that a small part of an infectious agent might serve as a vaccine goes back more than 45 years, to studies done

SYNTHETIC VACCINES are designed with the help of computer-graphics programs. These displays generated by Arthur J. Olson of the Research Institute of Scripps Clinic show a method whereby parts of a viral protein that are on the surface of a virus, and therefore accessible to antibodies, can be identified. The backbone of the surface domain of the protein on the outer shell of the tomato bushy-stunt virus is displayed (*1*) on the basis of coordinates determined by Stephen C. Harrison of Harvard University and his colleagues. A single peptide of the protein is picked out in yellow, with the side chains of its component amino acids indicated in atomic detail (*2*). The peptide is enlarged and a sphere representing a water molecule is displayed (*3*). The sphere is rolled around the peptide to generate a map of the surface accessible to water (*4*); it does so, following an algorithm developed by Michael L. Connolly, by placing a dot at each point of its closest contact with the peptide, taking account of the sphere's own van der Waals radius (zone of influence, in effect) and that of each atom of the peptide and the rest of the protein. A similar dot-surface map is generated to show what parts of the peptide are still accessible to water when three copies of the protein are associated in an array on the surface of the virus (*5*) and when four such arrays (out of 60) are in position on the outer surface of the virus (*6*).

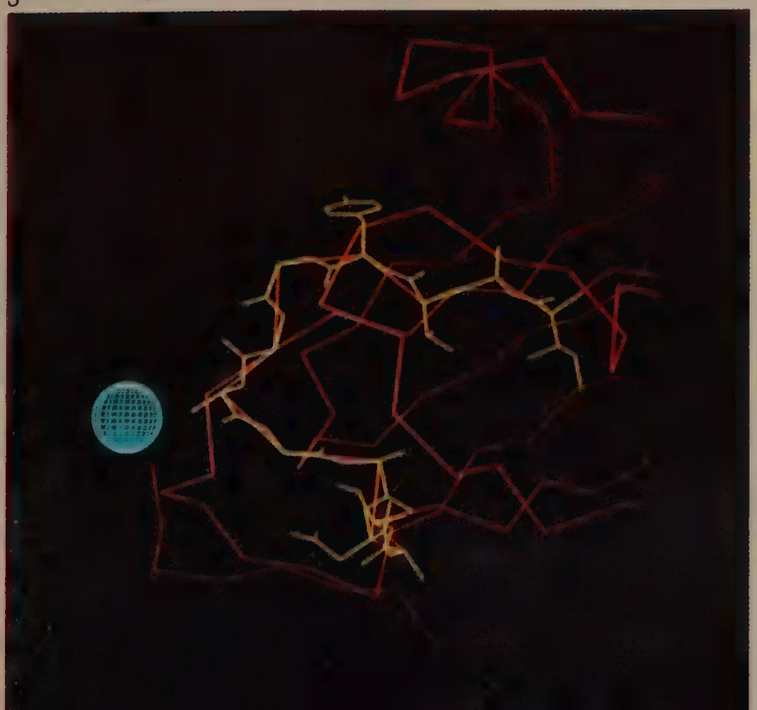
1



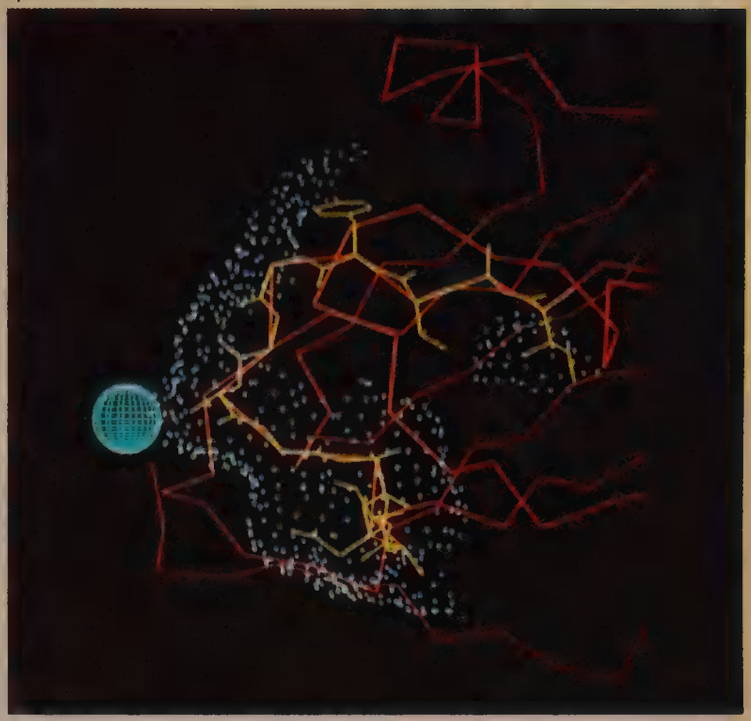
2



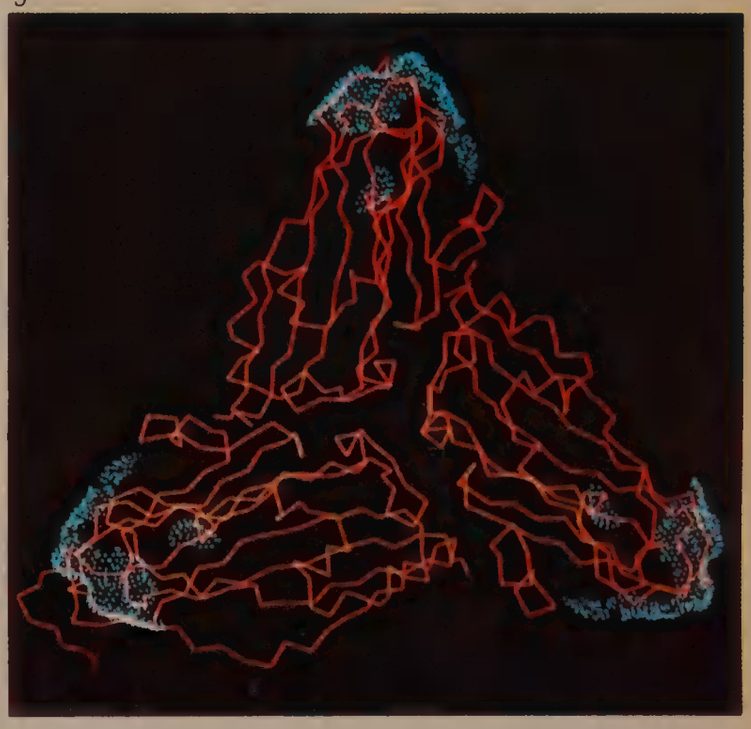
3



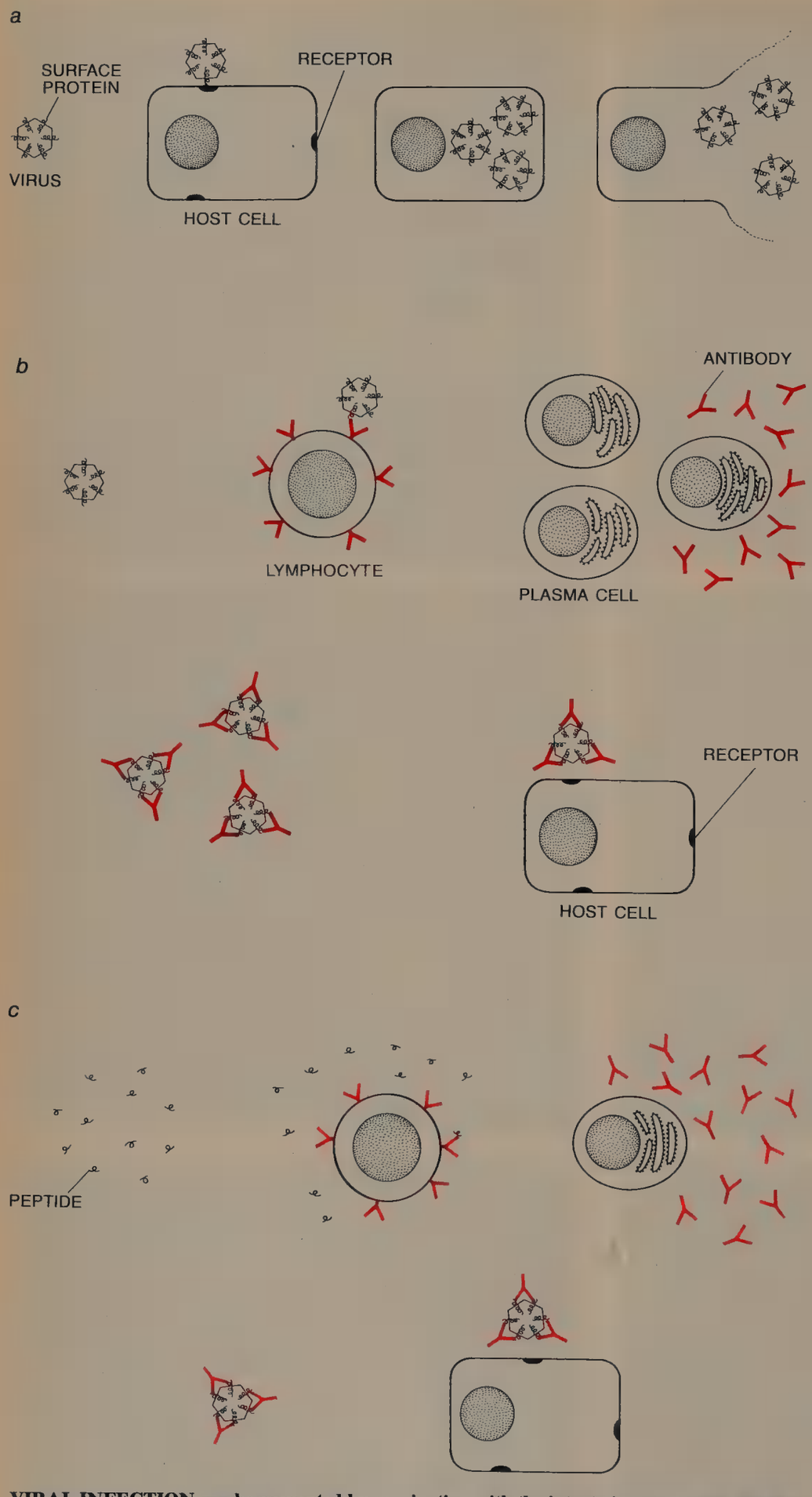
4



5



6



VIRAL INFECTION can be prevented by vaccination with the intact virus or a small peptide synthesized to mimic a site on a surface protein of the virus. A virus infects a cell (*a*) by attaching itself to a receptor on the cell, entering the cell and taking over the cell's machinery to make many copies of itself, which burst the cell. An attenuated or killed virus can be injected as a vaccine (*b*). When such a virus encounters a receptor on a lymphocyte that is genetically programmed to recognize and make antibody against a surface protein of the virus, the lymphocyte is stimulated to form a clone of antibody-secreting plasma cells. In any subsequent exposure to the virus the antibodies bind to the protein and neutralize the virus by blocking its attachment to host cells. A synthetic peptide mimicking part of the surface protein can serve as a vaccine (*c*). It too elicits antibodies able to bind to the surface protein and neutralize the virus.

by Walther F. Goebel of the Rockefeller Institute for Medical Research. He showed that immunizing mice with a sugar chain from the surface of the bacterium that causes pneumonia would protect the mice from subsequent infection. The feasibility of making immunogenic peptides to order was greatly increased in the 1960's when R. B. Merrifield of Rockefeller University developed a method for automatically linking amino acids, the subunits of proteins, to make a peptide.

It is not always easy to determine the amino acid sequence of a particular peptide in a large protein molecule, but a way around that difficulty was found. The amino acid sequence of a protein is specified by the sequence of the subunits called nucleotides in the nucleic acid (DNA or RNA) encoding the protein. A few years ago remarkable new techniques were developed for determining the nucleotide sequence of DNA. Now it became possible to very rapidly determine the complete sequence of the DNA of a virus (or of a DNA copy of a viral RNA). In principle, then, one could identify a nucleic acid sequence likely to encode a surface protein and translate that region according to the genetic code to derive the amino acid sequence of the protein. Then, exploiting Merrifield's procedure, one could synthesize short peptides in various regions of the protein, inject them in laboratory animals and see if they would give rise to antibodies that bind to the natural protein and perhaps neutralize the virus.

Before one could begin rationally designing vaccines, however, some difficult questions had to be answered having to do with the nature of an antigenic determinant and whether or not a short synthetic peptide could be expected to mimic one. Antibodies bind to protein sites by reason of shape and electric charge. The backbone of a protein is a chain of carbon and nitrogen atoms bristling with projections: the side chains of successive amino acids, each of which has a characteristic shape and array of charges. The sequence of the amino acids (which constitutes the primary structure of the protein) and their consequent short-range interactions determine how the chain twists locally into what is called the secondary structure: perhaps a helix or a sharp turn. Interactions among secondary structures in turn determine how the chain folds back on itself to form the complex tertiary structure characteristic of proteins.

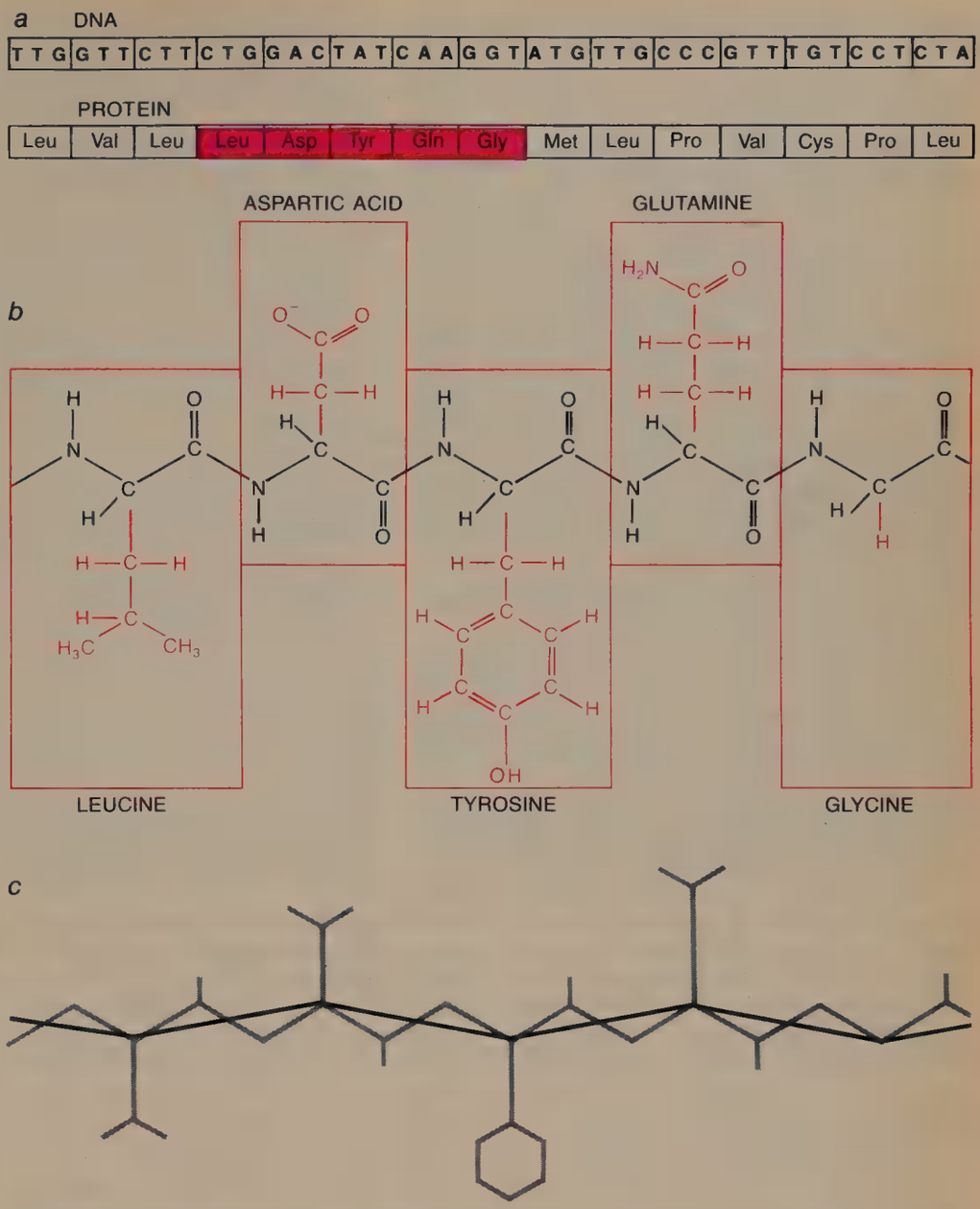
It follows that there can be two kinds of antigenic determinant. If the antigen-binding site of an antibody molecule recognizes and binds to an array of contiguous amino acid side chains, one speaks of a continuous antigenic determinant. A discontinuous determinant, on the other hand, is one made up of groups of amino acids that are separat-

ed from one another along the chain but are brought into proximity by tertiary folding.

As of two or three years ago a considerable body of immunological doctrine suggested that synthesized peptides were not likely to mimic the natural antigenic determinants of intact viral proteins. In the first place it appears that the immune response to a viral protein in the course of a natural infection is directed to a small number of antigenic determinants confined to a few parts of the protein molecule; specifically, there is evidence that a protein's immunogenicity (its ability to elicit antibodies) is accounted for by only one site for every 50 amino acids. Moreover, most of these relatively few antigenic determinants seem to be discontinuous ones. To make a peptide simulating such a determinant one would have to make a long stretch of the protein; such a synthesis is technically feasible but inefficient. If a long peptide were synthesized, it seemed unlikely to fold itself into the correct tertiary structure in the absence of the rest of the molecule. Indeed, it was thought the individual side chains in a short peptide would not spontaneously orient themselves to mimic even the simpler secondary structure of a continuous determinant.

To challenge these rather pessimistic assumptions we decided two years ago to synthesize in short segments almost the entire length of a viral protein and see whether antibodies to many of the peptides might not react with the entire protein. The molecule we chose was the hemagglutinin of one strain of the influenza virus. It is a glycoprotein (a protein with sugar chains attached) that forms spikes radiating from the spherical envelope of the virus. The hemagglutinin spikes have important functions. They mediate the binding of the virus to a host cell. As their name indicates, they cause red blood cells to clump together. And they are primarily responsible for eliciting antibodies that neutralize the virus.

Two recent accomplishments made the influenza hemagglutinin an attractive choice for our studies. Walter Fiers of the State University of Ghent in Belgium and his colleagues had determined the nucleotide sequence of the gene encoding the protein, and from it they had predicted the amino acid sequence. Ian A. Wilson and Don C. Wiley of Harvard University and John J. Skehel of the National Institute for Medical Research in London had determined the structure of the protein by X-ray crystallography, so that we would be able to relate each peptide we synthesized to a particular region of the secondary and tertiary structure. A molecule of the hemagglutinin is composed of two protein chains designated *HA*1 and *HA*2; each spike is a trimer consisting of three such molecules. *HA*1 seems to be the primary target of the immune system, and so we turned first to that chain.



NUCLEOTIDE SEQUENCE of a stretch of viral DNA encodes the amino acid sequence of a viral protein (*a*). Once the DNA sequence is known, investigators can predict the amino acid sequence and pick out for synthesis those peptides, or short segments of the protein chain, that seem likely to give rise to antibodies. The chemical structure of part of the peptide shown in a is diagrammed (*b*). The main chain is composed of carbon (C) and nitrogen (N) atoms with projecting hydrogen (H) and oxygen (O) atoms. Each amino acid is characterized by a side chain (*color*) that projects from what is called the alpha carbon. In the computer displays of proteins that illustrate this article either the alpha-carbon backbone is shown (*c*) in simplified form (*black*) or the chain is shown in atomic detail but with hydrogen atoms excluded (*gray*).

We synthesized 20 peptides accounting for some 75 percent of the length of the chain and sampling a wide variety of secondary and tertiary structures. We injected each of the peptides, bound to a large carrier protein, into rabbits and tested serum from each rabbit (by what is called an enzyme-linked immunosorbent assay) for the presence of antibodies that would bind to the injected peptide, to the hemagglutinin molecule and to the intact virus. All 20 of the peptides gave rise to antibodies against themselves; 15 of the antipeptide antibodies reacted with the protein molecule and 17 of them bound to the influenza virus. In other words, a peptide from almost any region of a viral protein can elicit an antibody that will recognize the entire protein, either in purified form or on the surface of the virus; the natural immunogenicity of a protein (the only kind that has been much studied up to now) is less than the potential immunogenicity of its parts. The only requirement seems to be that the synthesized peptide be on the surface of the folded protein, where the antibody can get at it.

There are a number of ways to estimate whether a given peptide is on the surface of a protein—or, to put it more exactly, whether a peptide (or part of it) is accessible to antibody molecules in solution in an animal's blood. One of the most precise ways is to program a computer to make a graphic display of the

solvent-accessible regions [*see illustration on page 67*]. First the peptide under study is displayed, complete with its branching side chains, in the conformation it assumes on the intact protein molecule. Then the computer rolls a sphere representing a water molecule around the peptide. The sphere penetrates where it can, given the van der Waals radius (loosely, the zone of influence in terms of electric charge) of the water molecule and of the atoms of the side chains. The sphere leaves a dot at each point of contact with the peptide, thus in effect drawing a map of the solvent-accessible regions of the peptide.

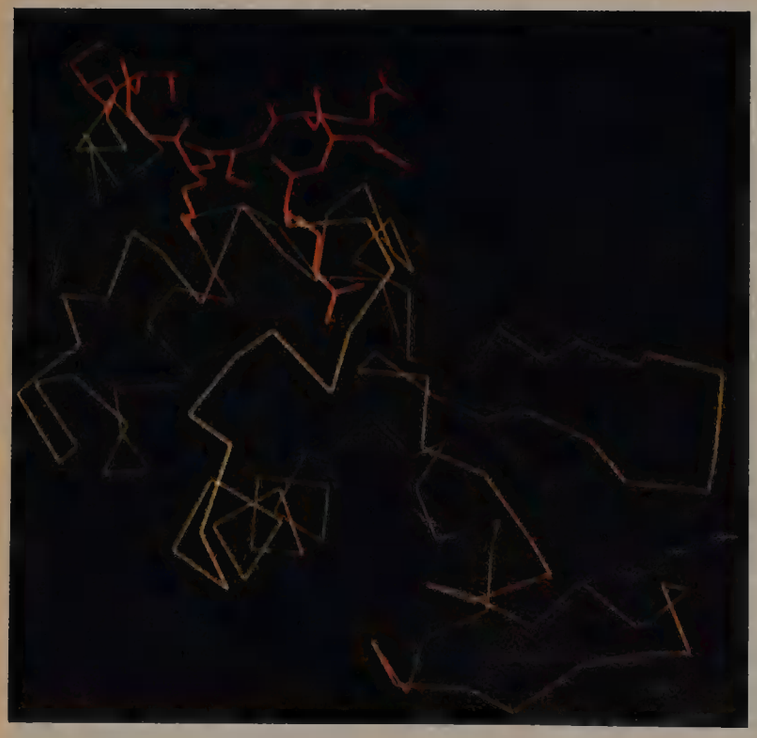
The scheme I have described is effective, but it does require that the structure of a virus and its proteins be known in atomic detail, and such data are not available for many viruses. Fortunately there is a way to predict which parts of a protein are likely to be on the surface if one knows only the amino acid sequence. The 20 amino acids can be characterized for the most part as being either hydrophilic (water-attracting) or hydrophobic (water-repelling). The hydrophobic amino acids are most often buried in the core of the folded protein, whereas the hydrophilic ones are usually on the outside of the molecule. The four most hydrophilic amino acids are lysine, arginine, aspartic acid and glutamic acid. One simply assumes the regions of a protein that are relatively rich in these amino acids are likely to be exposed to an aqueous environment and therefore are on the surface.

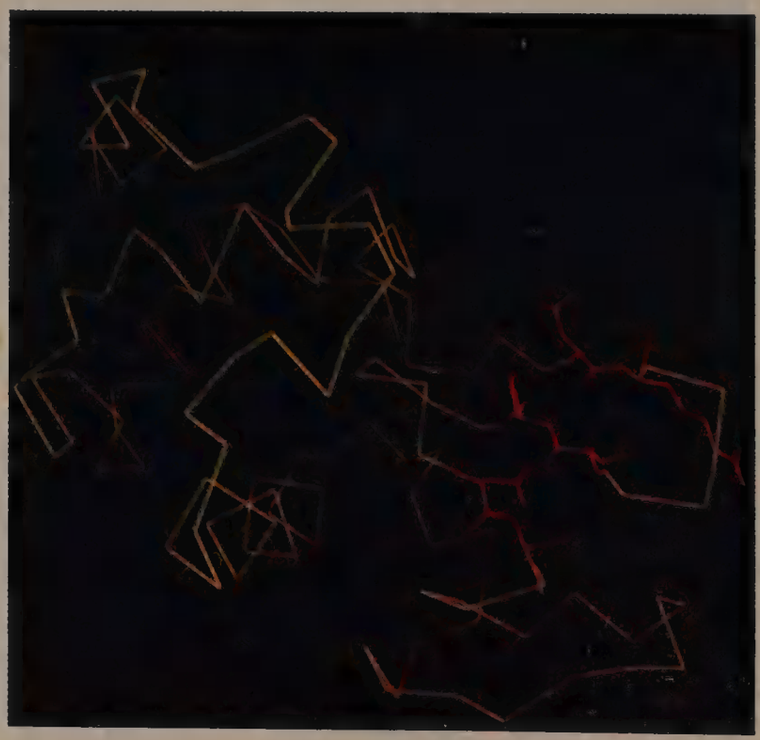
Our results with the influenza virus suggested, as I pointed out above, that just about any surface peptide should give rise to antibodies able to react with the entire protein. How is it that a lone peptide can be depended on to assume a conformation similar to its conformation when it is folded into the complete protein molecule? It may be the very lack of constraints on shape that gives rise to the correct shape. An injected small peptide, in solution in an animal's bloodstream, keeps twisting and bending. As it changes shape it presumably continues to encounter lymphocytes, and hence to trigger the secretion of antibodies against each shape in turn. One of those shapes, it now seems clear, is likely to coincide closely with the conformation the peptide adopts as part of a large protein. (It is interesting to note that in many cases small peptides are more effective in eliciting antibodies against a viral protein than the purified protein is. Perhaps the isolated protein locks itself into a folded configuration that is different from the shape it assumes when it is arrayed with many copies of itself on a virus.)

If virtually any surface region of a viral protein can elicit antibodies against that protein, the way is open for designing a vaccine rationally instead of injecting whole virus particles and in effect taking whatever the immune system offers. That is being done in a number of laboratories. The first vaccine we developed was against the virus of foot-and-mouth disease, a historic plague of animals (cattle in particular) that is still a serious problem on every continent except Australia and North America. Some three billion doses of killed-virus vaccine are administered annually in an effort to control it. The virus shell is composed of 240 protein molecules, 60 copies each of the four viral proteins *VP*1, *VP*2, *VP*3 and *VP*4. There have been attempts to make a vaccine based on *VP*1, which is thought to be the major natural target of neutralizing antibody, but the entire protein has proved to be not very immunogenic. In collaboration with Fred Brown and David J. Rowlands of the Animal Virus Research Institute in England, we undertook to synthesize a more immunogenic subregion of the protein.

The sequence of *VP*1's 213 amino acids had already been determined from the nucleotide sequence. The structure of the protein is not yet known in atomic detail. Analysis of the amino acid sequence, however, was enough to draw our attention to the region between amino acids 141 and 160. The region is fairly rich in the hydrophilic side chains of arginine, lysine and aspartic acid, suggesting it might be on the surface. The amino acid proline is often found where a protein makes a sharp bend (which may make for immunological specificity), and the region has two prolines.

Most important, it was known that in this region the amino acid sequence undergoes considerable variation from strain to strain. A virus needs to evade a host's immune-system defenses long enough to invade host cells and produce progeny. It does so by continually changing the antigenic determinants on its outer shell or envelope, with the result that antibodies formed earlier





ANTIGENIC DETERMINANTS are sites on a protein that are recognized by antibodies and to which the antibodies bind. The two kinds of determinants are illustrated by these computer displays, in which part of the main chain of the enzyme lysozyme is shown in yellow, with antigenic determinants (side chains included) picked out in orange. A continuous determinant (*left*) is one formed by an array of contiguous amino acids along one segment of the chain. A discontinuous determinant (*right*) is one composed of several segments that are separated from one another along the chain but are brought together to form a single site by the tertiary folding of the protein.

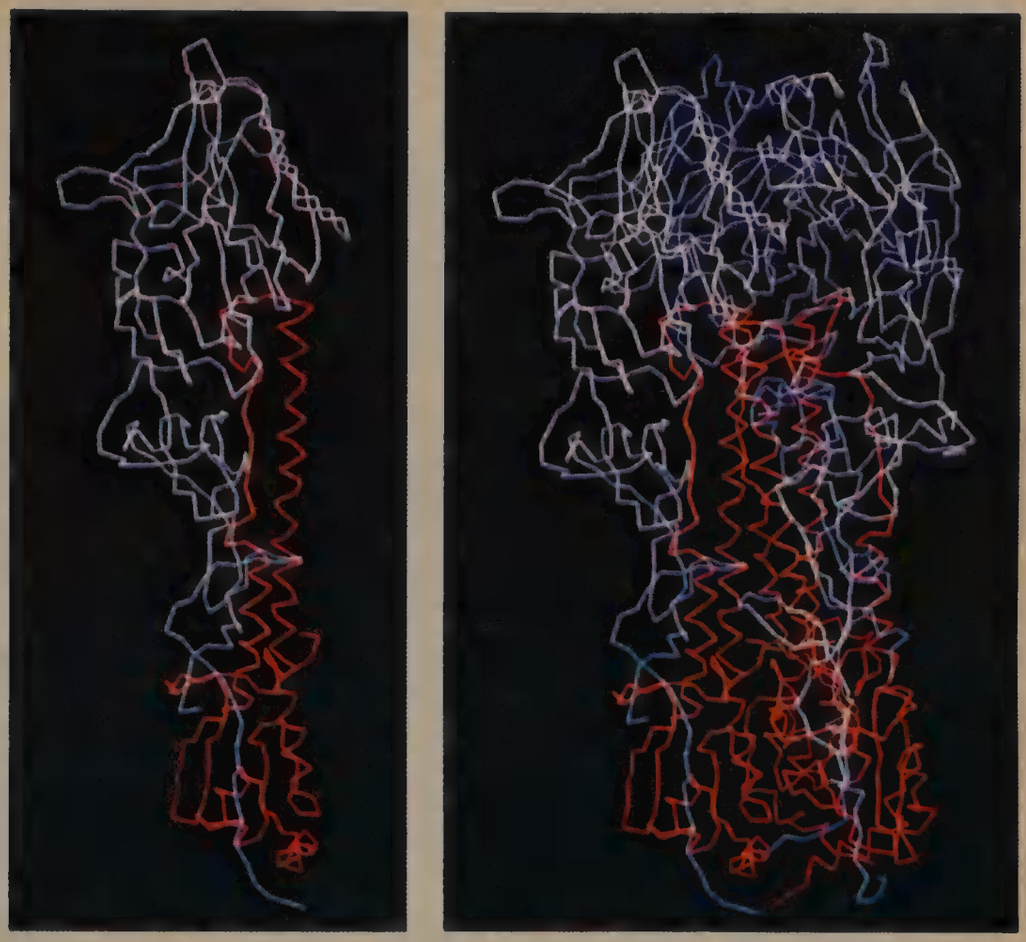
against a parental strain will not bind to the altered virus and neutralize it. In less teleological terms, what happens is that natural selection encourages the survival of a strain in which spontaneous mutation has brought about an immunologically significant variation. A region subject to such mutation is likely to be one that ordinarily elicits antibodies, and so a peptide from that region is a good candidate for a synthetic vaccine.

We synthesized a number of peptides from the *VP*1 of the foot-and-mouth virus, including the peptide spanning positions 141 to 160, coupled each one to a carrier molecule and injected it into rabbits. After determining that the peptides elicited antibodies against themselves, we did a bioassay to test their ability to neutralize the virus. When the rabbit antiserum to peptide 141–160 was injected into mice along with the virus, it effectively neutralized the virus; the mice were given what is called passive immunity by the injected antipeptide antibody. There are three major strains of the foot-and-mouth virus; we had worked with type *O*. When we tested the same antibody's effect on the other two strains, the neutralization was much weaker, supporting the impression that changes in this particular region of *VP*1 allow the virus to elude antibodies made against a different strain.

The real test of any vaccine is its ability to confer not only passive immunity but also active immunity, that is, to cause the immune system to synthesize its own antibodies for protection against a future attack. We immunized guinea pigs with one shot (200 micrograms) of the synthetic-peptide vaccine and after 35 days administered a powerful dose of the virus: 10,000 times the dose normally required to infect half of a group of animals. None of the immunized animals came down with the disease, whereas all the unimmunized control animals did. A similar active-immunity test in cattle is now under way in England. Already we have found that cattle immunized with the peptide secrete antibody at a concentration that should protect them from a virus challenge.

A practical vaccine will, of course, have to protect against all three strains of the virus. We have synthesized peptides from the same region (141 to 160) of the *VP*1 of type-*A* and type-*C* viruses, which have slightly different amino acid sequences. So far antibodies to these peptides have been bioassayed in mice, and they do neutralize viruses of type *A* and type *C* respectively. A synthetic three-strain vaccine against foot-and-mouth disease seems to be within reach.

Given that one can in principle complete the technological chain from nucleic acid sequence to synthetic-peptide vaccine, the next step is to explore certain aspects of vaccine design in more detail. Current research is aimed at finding additional and more sophisticated ways to pinpoint the regions of viral proteins that are most vulnerable to immunological attack and also ways to confer immunity of long duration.



HEMAGGLUTININ MOLECULE of the influenza virus is made up of two protein chains, *HA*1 (*purple*) and *HA*2 (*orange*). The three-dimensional structure of the molecule (*left*) is based on X-ray crystallographic data supplied by Ian A. Wilson and Don C. Wiley of Harvard University and John J. Skehel of the National Institute for Medical Research in London; only the backbone is displayed. Three *HA*1 chains and three *HA*2 chains are associated in the hemagglutinin trimer (*right*). Trimers form spikes on the spherical envelope of the virus particle.

There are essentially two kinds of region in viral proteins. In one kind of region variation is allowed; indeed, it is essential if the virus is to confound the host's immune system. A vaccine that gives rise to antibodies against such a region (*HA*1 of influenza or *VP*1 of foot-and-mouth disease, for example) can be expected to be immunogenic, but it must be tailored to cope with each variant strain. There are other regions in viral proteins that have a constant and essential biological function and whose structure cannot vary. The amino acid sequence of such a region is highly conserved among all strains of the virus. If an essential and conserved region is accessible to antibodies, a vaccine that elicits antibodies against it should be a universal vaccine, one from which the virus cannot escape.

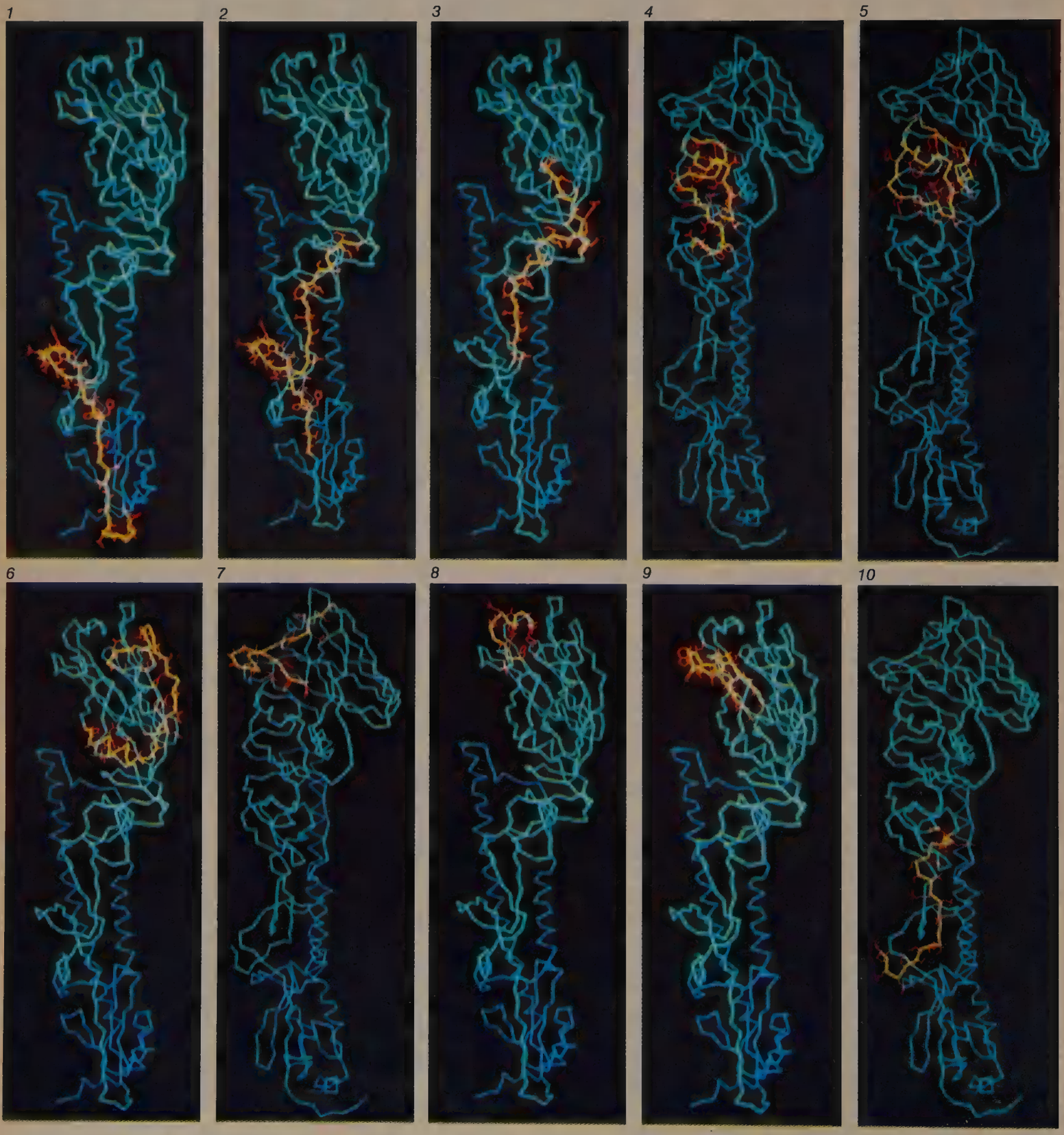
The influenza virus is notorious for its ability to evade the immune system by changing parts of the hemagglutinin on its surface. The changes are in the *HA*1 chain, and it was peptides from *HA*1 that we first tested for immunogenicity in the experiment I described above. The hemagglutinin has at least two other and unvarying functions, however. It forms a socket that binds to sialic acid, a sugar on the surface of cells, thereby initiating the infective process. It also causes the surface membranes of infected cells to fuse together, increasing viral infectivity. Evidence from a number of studies by Purnell W. Choppin of Rockefeller University and his colleagues and Michael Waterfield of the Imperial Cancer Research Fund suggests that a conserved region at the beginning of the *HA*2 chain is at least partly responsible for the fusion.

Normally the immune system does not "see" this region; antibodies to it are usually not made in the course of natural infection. If such antibodies were made (if the region were not only highly conserved and biologically essential but also naturally immunogenic), every human being and other animal host would long since have become immune to the influenza virus; there would be no influenza. We decided to see whether a synthetic peptide from the region might elicit antibodies that would neutralize more than one strain of the virus.

Stephen Alexander, Hannah Alexander and Nicola Green synthesized several peptides from the beginning of the

*HA*2 molecule; the amino acid sequence they chose was that of influenza type *A*, subtype H_3. Rabbit antiserums against these peptides effectively neutralized the *A*, H_3 virus when they were tested in a tissue-culture system, showing that even if this region of the *HA*2 molecule is not normally immunogenic, it is accessible to specific antipeptide antibodies. Moreover, the same antiserums neutralized viruses of a different subtype and even a different major type (influenza *B*). In contrast, antibodies elicited against the intact H_3 subtype of the virus did not neutralize other subtypes. Nor, of course, are individuals immunized by the injection of one influenza subtype thereby immunized against infection by a different subtype. Much more study of the anti-*HA*2 antibodies is needed, but these first results suggest that much may be expected of vaccines designed to attack a conserved region having an important biological function.

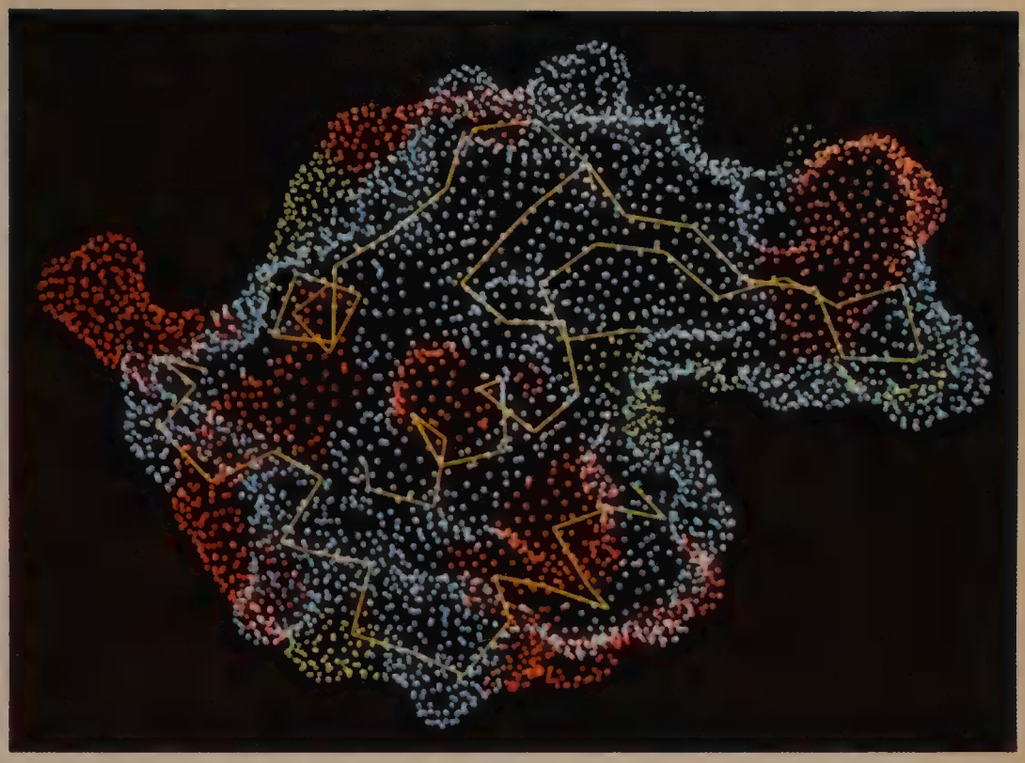
One way to design a vaccine, then, may be to choose for synthesis a region of a protein that can be expected, on theoretical grounds, to serve as an antigenic determinant. There is another way to proceed. One can try to mimic antigenic determinants that have already been shown to be important by seroepidemiologic studies, in which



SERIES OF PEPTIDES accounting for some 75 percent of the *HA*1 chain was synthesized and each peptide was tested for its ability to elicit antibodies reacting with the peptide, with the hemagglutinin molecule as a whole and with the influenza virus particle. Ten of the 20 peptides tested are shown in yellow in these displays, with their side chains in orange, the remainder of the *HA*1 chain in light blue and the *HA*2 chain in dark blue; in each case the hemagglutinin monomer has been rotated to show the peptide in question to best advantage. Most of the peptides elicited rabbit antibodies that reacted with the virus, supporting the concept of synthetic-peptide vaccines.

closely related viruses are classified on the basis of the antibodies they elicit. Such studies reveal the existence of different types and subtypes of a virus, such as those I have alluded to in discussing foot-and-mouth disease and influenza.

The types are defined by long-term studies in which viruses responsible for a disease are collected from various areas and at different times. Antibodies taken from patients are tested against the panel of viruses; different antibodies react with different strains of the virus because they "see" different antigenic determinants. By testing many antibodies against a panel of viruses serologists can describe a given virus type in terms of its pattern of determinants and can learn which determinants have been operative in a particular individual or epidemic. Until recently, however, the determinants were abstract "markers" designated by letters or numbers; they were not understood in molecular terms. It was therefore not at all clear that the vast amount of accumulated seroepidemiologic data could help in the design of vaccines. Synthetic peptides might not mimic the structures responsible for serologic markers, for example, if those structures were discontinuous in nature. We decided to try to design a vaccine against the liver disease hepatitis *B* by synthesizing peptides that would be the molecular equivalent of abstract serologic determinants that have been shown to be important.



SOLVENT-ACCESSIBLE SURFACE of the lysozyme molecule is mapped by the method demonstrated in the illustration on page 67, but with the dots color-coded according to the amino acid responsible for each part of the surface: orange for the basic amino acids lysine and arginine, green for the acidic ones aspartic acid and glutamic acid and blue for all others. The basic and acidic amino acids are the most hydrophilic (water-attracting) ones. As the display demonstrates, they tend to be exposed on the surface. If a protein's structure is not known in detail, one can simply assume peptides rich in hydrophilic amino acids will be immunogenic.

The viral component that appears to be the major target of neutralizing antibody, the hepatitis-*B* surface antigen, is a protein on the envelope of the virus particle. Serologic studies show that the several strains of the hepatitis-*B* virus have in common one determinant, which is designated *a*. Each strain also has two other determinants: either *d* or *y* and either *w* or *r*. That means there are four possible types of the virus: *adw, ayw, adr* and *ayr*. Working with John L. Gerin of the Georgetown University School of Medicine and Robert H. Purcell of the National Institute of Allergy and Infectious Diseases, we wanted to synthesize peptides that would specifically elicit antibodies against the *d* and the *y* determinants. The hepatitis-*B* surface antigen, however, is a chain of 226 amino acids. We needed a guide to where to begin. Analysis of the antigen's amino acid sequence in three strains of the virus by several investigators had shown that there are substantial differences among the strains in only one region of the protein, between amino acids 110 and 140. Any difference in determinants must originate with differences in amino acid sequences, and so that region was clearly the place to seek our peptides.

Following one of the three available amino acid sequences, we synthesized several peptides spanning the region and made antibody to each of them. One 13–amino acid peptide (positions 125 to 137) elicited antibody that reacted only with viruses bearing the *y* determinant. This meant that the *y* determinant and its alternative, the *d* determinant, must lie within the 13–amino acid peptide. Having synthesized *y*, we should be able to synthesize *d* by following a different one of the other known sequences for that region. The sequence we chose next differs from the first one at just three positions: two threonines and a tyrosine are replaced respectively by proline, asparagine and phenylalanine. Antibody to the new peptide reacted only with viruses bearing the *d* determinant.

Perhaps the switch from *y* to *d* is brought about by fewer than three amino acid substitutions. By synthesizing peptides in which only one of the three positions is changed at a time we should soon learn whether one, two or all three of the substitutions are required. So far the experiment has taught us two things. One is that very subtle changes in the amino acid composition of a protein can define different strains of a virus. The other is that synthetic peptides can mimic the distinctions revealed by serologic studies; in designing synthetic vaccines one will be able to take advantage of serologic evidence and thereby, for example, attempt to cope with a current epidemic occasioned by a mutant strain.

Before synthetic vaccines can become an alternative to conventional ones ways must be found to increase the magnitude and duration of the immunity they confer. An advantage of attenuated viruses, as I have mentioned, is that they keep proliferating and thus present ever more antigen to the immune system, thereby increasing the antibody level. A synthetic peptide, like a killed virus, does not replicate, and so substances called adjuvants must be added to the vaccine to increase the level and duration of immunity.

One of the most powerful of these, Freund's adjuvant, is an emulsion of mineral oil and water mixed with killed bacteria (mycobacteria, a class that includes the agent of tuberculosis). Just how it works is not known, but it probably does two things. The emulsion retains antigen trapped in it, forming a depot from which antigen is released slowly, simulating the sustained presentation of antigen by a replicating virus. Meanwhile the killed bacteria attract to the depot various cells that have roles in the immune response. To attract such cells is also to serve as an irritant, however, causing soreness at the site of injection and even the formation of an abscess. Powerful adjuvants such as Freund's may not be administered to human beings; instead, alum, or aluminum hydroxide, is commonly included with a killed-virus vaccine. The antigen is adsorbed onto the aluminum hydroxide particles and is released slowly after injection. In some cases, it appears, alum

will serve well enough with synthetic vaccines, but most such vaccines will require something better. Much effort is currently aimed at the development of safe and effective adjuvants.

Some hopeful results in this direction have been reported by François Audibert and Louis Chedid of the Pasteur Institute and Ruth Arnon and Michael Sela of the Weizmann Institute of Science, who recently elicited antibodies conferring passive immunity against the diphtheria toxin. They synthesized three peptides corresponding to regions of the toxin molecule. Before injection a synthetic peptide must be coupled to a carrier molecule. (In all our experiments the carrier was keyhole-limpet hemocyanin, a respiratory pigment of a mollusk, which is often used for this purpose.) The Pasteur-Weizmann group devised a carrier-adjuvant combination. They coupled their diphtheria-toxin peptides to a carrier that had been linked to a simple derivative of the cell membrane of mycobacteria; the membrane derivative apparently serves as an adjuvant, significantly increasing the immunogenicity of the peptides.

SYNTHETIC PEPTIDE is injected into a rabbit (*1*) to test its ability to elicit antibodies. Serum withdrawn from the rabbit (*2*) can first be assayed for the presence of antibodies able to bind to the injected peptide (*3*), to the viral protein in a whole (*4*) and to the virus itself (*5*). The binding of the antipeptide antibody (*color*) is detected by noting the activity of an enzyme coupled to a second antibody (*black*) that binds to the first one. The ability of the antipeptide antibody to neutralize the virus (to confer passive immunity) is tested by injecting mice with the virus and the antiserum (*6*); if the mouse stays well, the virus has been neutralized. The ability of the synthetic peptide to serve as a vaccine conferring active immunity can be tested by injecting the peptide into a guinea pig (*7*) and later injecting the virus (*8*). If animal survives (whereas unimmunized control animals do not), efficacy of vaccine is demonstrated.

As I mentioned at the beginning of this article, antipeptide antibodies should have important applications as investigative tools. One application arises from advances in nucleic acid technology. Because of the precision of recombinant-DNA techniques and the remarkable speed of the new methods for determining the sequence of a nucleic acid, it is often the case that one has the complete sequence of a gene before the protein it encodes is known; one has a gene in search of a protein. From the DNA or RNA sequence one can now predict an amino acid sequence and select a small region of it for synthesis. Antibodies elicited by the peptide serve as a probe for finding the protein. By this means, for example, the proteins encoded by some oncogenes, the tumor-inducing genes of certain viruses, have been identified and tracked to particular sites in cells.

The distinguishing feature of antipeptide antibodies is of course their predetermined specificity. They are antibodies not to a protein but to a particular part of the protein that is known to the experimenter. Once they have been manufactured by the immune system of a laboratory animal they can easily be purified by a single passage through an immunoadsorbent column. Then they can be used to explore in detail the structure of a protein and to correlate the structure with function. For example, antibodies to a particular part of a protein will bind to that part and inhibit its function, and so tests with a battery of peptides should make it possible to define the active site of a protein or various sites that are active in different functions. Again, many peptide hormones originate in a long protein chain that is subsequently cleaved by an enzyme into smaller pieces, each one a hormone with a particular activity. Antibodies to peptides serve to sort out the cleaved pieces and correlate each with its activity.

Finally, antibodies of predetermined specificity can reveal just which part of a gene is being expressed to make a protein at a particular time. Mammalian genes are split into a number of the coding regions called exons. In the case of certain variable proteins different exons are expressed at different times. Successive forms of antibody molecules are synthesized, for example, in the course of the immune response. Now one can synthesize the peptide encoded by each exon, make antibody to it and learn which exon is being expressed at a particular stage of the immune response.

1700 pounds of optical perfection.

Perfection is not a term we use lightly. Or frequently. We've been taught for generations that perfection, in lens-optics terms, means ¼ wavelength. We've bettered this tolerance in the manufacture of an "egg crate" mirror, one-fifth the weight of a solid mirror. Manufactured for the Space Telescope Program, every point on its nearly 8-foot-diameter aspheric surface is within ⅒ wavelength of perfection. So while we're not 100% perfect, we're within several millionths of an inch of that goal.

We put all our experience into play in the manufacture of this massive optical element 97½ inches in diameter and 1700 pounds in weight. And our challenges were many. Since this lens is intended to perform in orbit, where the pull of gravity has been neutralized, we had to solve a gravity-distortion problem during the manufacturing cycle. We solved the problem with an air cushion, a membrane that inflates for testing to provide a uniform support and thus eliminate gravity-induced strain. The support deflates for polishing operations.

An aspheric mirror to gather information from the universe is an example of our large-optics capability. (Optics of this kind are ground, polished, and tested in our own lab with its vibration-isolated, six-story-high interferometric test towers.) But we're big in small things, too. Like lenses for cameras and copiers. We meet customers' needs with pride, precision, and stringent quality control.

In addition to massive, sophisticated equipment, a century of optical progress, and systems utilizing our ability to the fullest, we offer customers a tradition of realism when predicting costs and delivery.

That's important. Whether you seek to get a clearer picture of your neighbor or of celestial objects.

For more information about Kodak's contribution to the Space Telescope Program, write for the booklet *Mirror for the World's Most Powerful Telescope.* Send your request to: Richard M. Price, Research and Engineering, Kodak Apparatus Division, 901 Elmgrove Road, Rochester, NY 14650.

Kodak

Kodak. Where technology anticipates need.

© Eastman Kodak Company, 1983

SCIENCE AND THE CITIZEN

Early Warning

The apparent demise of the Reagan Administration's plan for deploying 100 MX missiles in a closely spaced array of "superhardened" missile silos has left strategic military planners in a quandary. The "dense pack" scheme was the latest of more than 30 basing modes for the new missile that have been considered seriously at one time or another by successive administrations; on close scrutiny all of them have been found wanting for a variety of technical, economic and political reasons. At this stage it is beginning to seem to many people, both in and out of the Government, that there simply is no basing mode for the MX that would guarantee its ability to ride out a nuclear attack by increasingly accurate Russian land-based missiles.

Such a conclusion is likely to focus renewed attention on yet another solution to the perceived problem of the vulnerability of the U.S. land-based missile force: the idea is to install the MX missiles in existing Minuteman missile silos and to adopt a "launch on warning" strategy. U.S. surveillance satellites are thought to be capable of providing a 30-minute warning of an attack by Russian land-based missiles aimed at the Minuteman sites. In that interval the U.S. missiles (either Minuteman or MX) could be launched in the opposite direction, leaving only empty holes as targets for the incoming Russian missiles. One attraction of such an approach is economic: it would not require the construction of a costly new basing system for the new generation of missiles. The chief obstacle is political: many Americans would presumably object to the automatic nature of such a response.

The present situation was not entirely unforeseen. Critics of the U.S. role in the nuclear-arms race have argued for some time that the land-based missiles of both the U.S. and the U.S.S.R. would eventually become vulnerable to attack and that greater efforts should therefore be made to phase them out of existence in favor of less vulnerable, and hence less destabilizing, submarine- and aircraft-based strategic weapons. Writing in these pages three administrations ago, Herbert F. York, former director of research and engineering for the Department of Defense, examined the various alternatives that had been advanced by then for protecting land-based missiles. The proposals included the deployment of antimissile defenses, the superhardening of silos, the proliferation of land-based missiles, a "shell game" in which there are more silos than missiles and the use of mobile missiles. York found that none of the plans "holds much promise beyond the next 10 years" (see "Military Technology and National Security," by Herbert F. York; SCIENTIFIC AMERICAN, August, 1969).

According to York, "one and only one technically viable solution seems to have emerged for the long run: Launch on warning." To most people, including him, he added, "such an approach to the problem is politically and morally unacceptable, and if it really is the only approach, then clearly we have been considering the wrong problem" in attempting to preserve the land-based component of the U.S. strategic deterrent force. For the time being, he concluded, submarine- and aircraft-based strategic weapons would appear to provide a satisfactory alternative to land-based missiles. "I expect, however, that as the continuing national debate subjects the whole matter of strategic arms to further public scrutiny we shall learn that these other alternatives also have dangerous flaws, and we shall see confirmed the idea that there is no technical solution to the dilemma of the steady decrease in our national security that has... accompanied the steady increase in our military power."

The Inflationary Universe

According to the standard big-bang cosmology, the universe began as a point of infinite density and expanded to a volume about the size of a softball by the end of the first 10^{-35} second. At this stage the universe was already far too big and had expanded too quickly to have yet been crossed by a signal moving at the speed of light. Causal connections between events, which are limited to the speed of light, were confined to regions about 10^{-25} centimeter across; some 10^{78} such regions would fit in a universe the size of a softball.

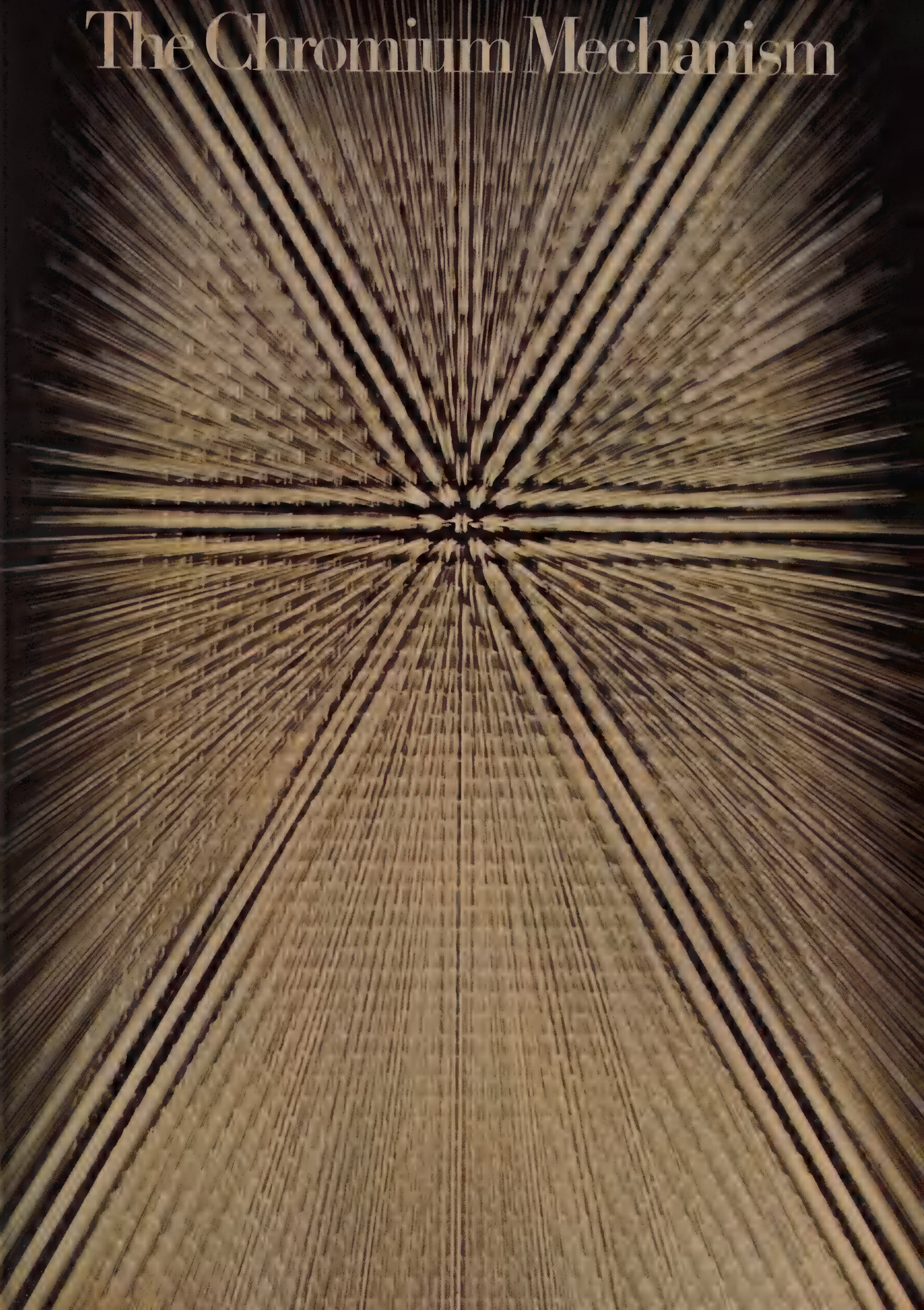
Physical theory now makes it possible to solve several outstanding problems in cosmology by supposing that the universe observable now was once an extraordinarily minute part of such a causally connected region. For a period beginning 10^{-35} second after the big bang and continuing for 10^{-32} second the region may have passed through what is called the inflationary phase. The region blew up by perhaps 100 orders of magnitude, from a trillionth the diameter of a proton to a diameter 10^{47} times that of the current observable universe. (The precise inflation factor is not important as long as it is greater than 10^{25}.) At the end of this titanic expansion the region that was to become the current observable universe must have been the size of a softball, embedded in an unimaginably greater universe that has ever since continued to expand beyond our horizon of cause and effect. At age 10^{-35} plus 10^{-32} second the inflationary phase was replaced by the standard phase, and the softball-size region, as well as the surrounding universe, resumed the rate of expansion that has since governed cosmic evolution.

These striking proposals have been made by a group of physicists and cosmologists who see in the physics of elementary particles a path to a consistent and perhaps even a predictive understanding of what the universe was like only 10^{-35} second after it began. The idea originated in a paper published last year in *Physical Review* by Alan H. Guth of the Massachusetts Institute of Technology. Guth noted certain inconsistencies of the model in his paper, but within a year Andreas Albrecht and Paul Steinhardt of the University of Pennsylvania and, independently, A. D. Linde of the Lebedev Physical Institute in Moscow were able to patch up some of the difficulties. Now these investigators and several others have taken up revisions of the inflationary model and are pursuing its consequences.

Guth's central idea was to apply thermodynamic principles to the interactions of matter and energy in a universe whose temperature is approximately 10^{27} degrees Kelvin. At this enormous temperature a typical particle has an energy at which the distinctions between three of the fundamental forces of nature are thought to break down. At ordinary temperatures there are four such forces: the strong force, which is responsible for nuclear interactions; the weak force, which accounts for the decay of the neutron into a proton; the electromagnetic force, and the gravitational force. One of the most ardent hopes of theoretical physics has been to give a unified account of all four forces, but so far only the first three have been linked by the grand unified theories. Above the critical temperature of 10^{27} degrees K. the theories predict that the strong, the weak and the electromagnetic forces are symmetrical, in the sense that they are indistinguishable from one another. Below the critical temperature the strong force should act differently from the other two forces, and the symmetry is broken. One effect of symmetry breaking, Guth suggests, is that matter and energy enter a new phase.

The phase transition envisioned by Guth is modeled closely on the conversion of a liquid into a solid. The liquid-to-solid transition is accompanied by a rearrangement of the molecules of the liquid into a regular crystalline array. The rearrangement lowers the total en-

The Chromium Mechanism



The Chromium Mechanism

The first comprehensive explanation of electrochemical activity during the plating of chromium has recently been formulated at the General Motors Research Laboratories. This understanding has aided in transforming chromium plating into a highly efficient, high-speed operation.

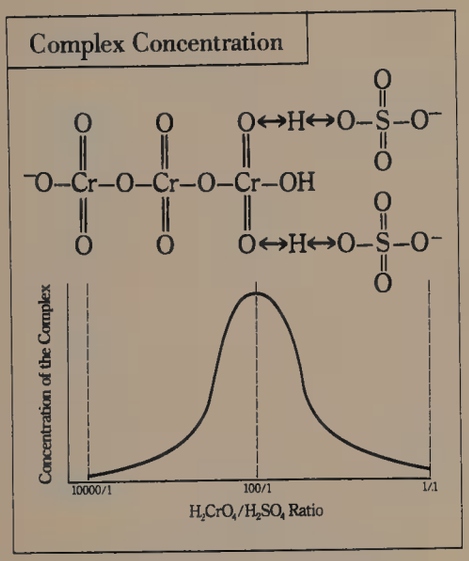


Figure 1: The electroactive complex and a theoretical plot of its concentration as a function of chromic acid to sulfuric acid ratio.

Figure 2: The electroactive complex diffuses from the bulk electrolyte solution (A) through the diffusion layer (B) to the Helmholtz double layer (C) to be discharged as metallic chromium (D) on the cathode (E) surface.

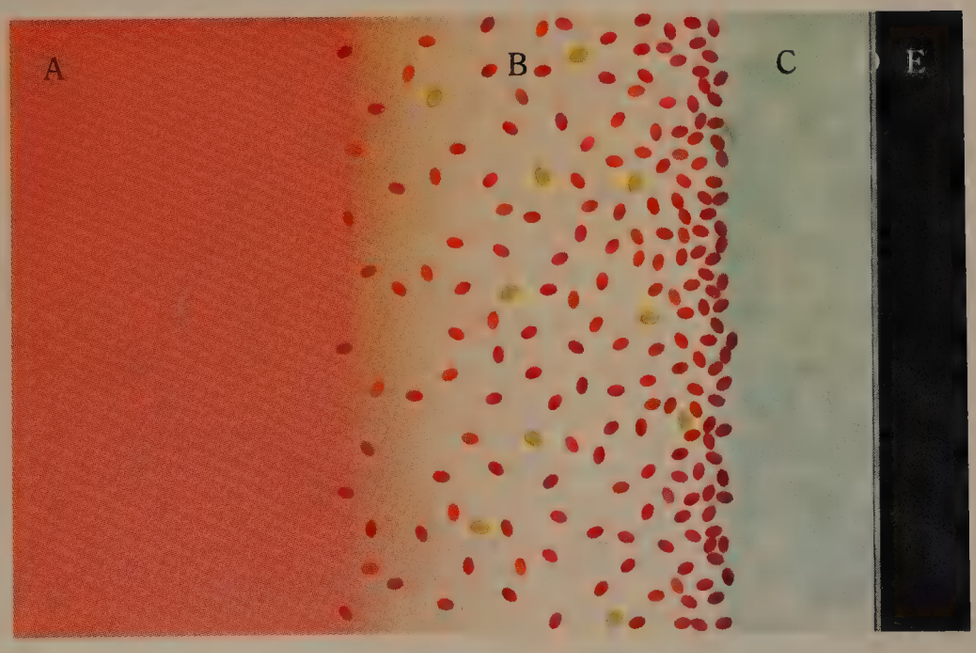
FOR MANY industrial applications, chromium coatings of more than 0.2 mil thickness are required for wear and corrosion resistance. But the conventional method of plating chromium is neither fast nor efficient. Nor, until the recent work of a GM researcher, had the steps involved in the century-old plating process been explained in detail. Through a combination of theory and experiment, Dr. James Hoare has devised the first comprehensive mechanism for chromium plating. This increased understanding has helped electrochemists at the General Motors Research Laboratories develop a system that plates chromium sixty times faster than the conventional method, while improving energy-efficiency by a factor of three.

The electrolyte for plating is a chromic acid solution which contains various chromate ions: chromate, dichromate and trichromate. From a series of steady-state polarization experiments, Dr. Hoare concluded that trichromate is the ion important in chromium deposition.

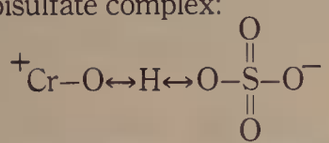
Sulfuric acid has been recognized as essential to chromium plating and has been assumed by some to be a catalyst for the process. In this strongly acidic solution, sulfate should be mostly present as the bisulfate ion $(HSO_4)^-$. Dr. Hoare found, contrary to expectations, that the addition of sulfuric acid to the plating bath decreased the conductivity of the solution.

Combining these findings with the results of previous investigations, Dr. Hoare concluded that the electroactive species was a trichromate-bisulfate complex (see Figure 1). From equilibrium considerations, he theorized that the maximum concentration of this species occurred at a 100-to-1 chromic acid/sulfuric acid ratio. The observation that the maximum rate of chromium deposition also occurred at this ratio supports the conclusion that this trichromate-bisulfate complex is the electroactive species.

During the plating process, the complex diffuses from the bulk solution toward the cathode (see Figure 2). Electron transport takes place by quantum mechanical tunneling through the potential energy barrier of the Helmholtz double layer and the unprotected chromium in the complex (Cr atom



on the left in Figure 1) loses electons by successive steps, going from Cr^{+6} to Cr^{+2}. Decomposition of the resulting chromous dichromate complex takes place by acid hydrolysis to form a chromous-oxybisulfate complex:



The positive end of this complex is adsorbed onto the cathode surface. Electrons are transferred from the cathode to the adsorbed chromium ion, forming metallic chromium and regenerating the $(HSO_4)^-$ ion. Thus, Dr. Hoare's mechanism explains how sulfuric acid, in the form of the bisulfate ion, participates in the plating process.

IT HAS long been known that chromium cannot be plated from a solution when initially present as Cr^{+3} because of the formation of the stable aquo complex, $[Cr(H_2O)_6]^{+3}$. Yet chromium can be plated when initially present as Cr^{+6} even though it must pass through the Cr^{+3} state before being deposited. Dr. Hoare's mechanism handles this paradox by explaining that the chromium ion being deposited (on the left in Figure 1) is protected by the rest of the complex as it passes through the Cr^{+3} state, so that the stable aquo complex cannot form.

The diffusion of the electroactive complex apparently controls the rate of the process, so that shortening the diffusion path increases the speed of chromium deposition. A high rate of relative motion between the electrolyte and the cathode will shorten the path. This can be accomplished by rapid flow or by agitation of the electrolyte.

Dr. Hoare found that the rate of chromium deposition increased with electrolyte flow until the process was no longer diffusion-controlled. He also found that the use of dilute electrolyte significantly increased plating efficiency.

"This project is an excellent example," says Dr. Hoare, "of how basic research and engineering principles can be combined to develop a new, successful process. Now, we'd like to take on the challenge of plating successfully from Cr^{+3}, which would be an even more efficient way to provide corrosion and wear resistance."

General Motors

GM

MARK OF EXCELLENCE

THE MAN BEHIND THE WORK

Dr. James Hoare is a Research Fellow at the General Motors Research Laboratories. He is a member of the Electrochemistry Department.

Dr. Hoare served as an electronics technician in the U.S. Navy during the Second World War. In 1949, he received his Ph.D. in physical chemistry from the Catholic University of America. After an assistant professorship at Trinity College in Washington, D.C., he joined the US Naval Research Laboratory as a physical chemist. He became a staff member at General Motors in 1960.

Dr. Hoare's sustaining interest has been in electrochemical kinetics and the mechanisms of electrode processes. He is best known to the scientific community for his basic studies of hydrogen and oxygen electrode mechanisms. His book, *The Electrochemistry of Oxygen,* published in 1968, is considered a work of primary importance to the field. In addition to his work on chromium plating, he is responsible for the fundamental research that helped make electrochemical machining a precision process.

ergy of the system, and so during the phase transition a certain quantity of energy called latent heat must be removed. It is possible, however, to cool a liquid below its normal freezing point if there are no sites that can serve as nuclei for the formation of crystals. The result is a new phase called a supercooled liquid; on contact with suitable nuclei such a liquid freezes but the release of latent heat can raise the temperature of the resulting solid.

In a similar way, Guth maintains, the symmetrical phase of matter and energy predicted by grand unified theories may have supercooled during the early stages in the evolution of the universe. The symmetrical phase might have been maintained in certain regions of the early universe at temperatures as low as 10^{22} degrees K. In the supercooled symmetrical phase matter and energy expand exponentially until the regions of symmetrical phase are converted into an asymmetrical phase, a relic of which we now perceive throughout the visible universe. During the phase transition the expanded region of symmetrical phase releases a tremendous quantity of latent heat, and the region is reheated to nearly 10^{27} degrees.

Although there remain several theoretical problems for the inflationary model, the model resolves at least three cosmological puzzles. One of them is the homogeneity of the cosmic microwave background radiation, which varies with direction by less than one part in 10,000. If the standard big-bang model is correct, regions of the sky separated by more than two degrees of arc have not been in contact throughout the entire history of the universe. No causal influence would have had time to propagate between them. Because the uniformity of the radiation cannot be accounted for by any physical interaction between distant regions, it must be introduced as an ad hoc hypothesis in the standard model. In the inflationary model such homogeneity is readily explained because the precursor of the present observable universe is much smaller than the one predicted by the standard model. Every part of the observable universe was at one time causally connected to every other part.

The second puzzle is called the flatness problem. A basic question in cosmology is whether the universe is open or closed: will it continue expanding forever, or is the density of matter sufficiently high for gravitational forces to halt the expansion and drag the universe back into a single point? If the density is exactly equal to the critical density that allows the expansion to continue forever, the universe is said to be flat: the geometry of space-time is not curved but instead approaches the Euclidean geometry of a plane.

The actual density of the universe has not been determined precisely, but it is known to be within one or two orders of magnitude of the critical density. For understanding the evolution of the early universe, however, the problem is not whether the present universe is open or closed but rather how the observed density came to be so close to the critical value. In the standard big-bang model the density of the early universe must be equal to the critical density to an accuracy of one part in 10^{49} in order to be consistent with the present range of values. In the inflationary model such precision is not necessary. During the exponential expansion of the universe the region occupied by the observable universe becomes increasingly flat, and the mass density naturally converges to the critical density. The observed universe is flat for much the same reason that the geometry of a small patch on the surface of a balloon approximates the geometry of a plane.

The third puzzle is a rather technical one: How can the production of the exotic particles called magnetic monopoles, which the grand unified theories say ought to exist, be reconciled with the big bang? Monopoles are generated during the phase transition from the symmetrical to the asymmetrical phase, and the number of monopoles generated is proportional to the rate of the phase transition. In the standard big-bang model the transition is so fast that monopoles are produced copiously; indeed, there are enough to imply an unreasonably high density of matter in the universe. In the inflationary model the phase transition is slower and the monopoles are not nearly as abundant.

The most serious difficulty for the inflationary model is accounting for the clumps of matter that are the most salient feature of the observable universe: galaxies, stars and planets. Small fluctuations in the density of the early universe either disappear entirely or collapse far too quickly to account for the observed distribution of matter. Galactic clumping, in the view of several investigators, may have come about later.

Autosplicer

In virtually all organisms genetic information is encoded in the nucleic acid DNA and is transcribed into three species of the similar nucleic acid RNA: messenger RNA, which is translated into protein, and ribosomal RNA and transfer RNA, which have roles in the process of translation. The various steps in transcription and translation are mediated by enzymes, the protein catalysts of biochemical reactions. There are enzymes called nucleases, for example, that cleave a nucleic acid into pieces and ligases that connect the pieces or form a single piece into a circular molecule. It has been taken for granted that for every such reaction there is an enzyme; nucleic acids do not cleave and ligate and circularize themselves by their own bootstraps. Or do they? Workers at the University of Colorado at Boulder have now found firm evidence that in the one-celled protozoan *Tetrahymena thermophila* there is an RNA that does just that.

In several species of *Tetrahymena* the genes coding for ribosomal RNA are split, that is, the coding region is interrupted by an intervening sequence (or "intron"). The entire gene is transcribed into a precursor RNA, which is then spliced: the intervening sequence is cut out of the precursor and the two long flanking sequences ("exons") it had interrupted are ligated, or joined by a covalent chemical bond, to form the mature ribosomal RNA. So far so good: since 1977 it has been known that in organisms higher than bacteria such intervening sequences interrupt many genes for messenger, transfer and ribosomal RNA. When Thomas R. Cech and his colleagues at Colorado studied the splicing process in *T. thermophila,* however, they found some surprises. One surprise was that the intervening sequence excised from the precursor RNA as a linear molecule is soon converted into a circular form. Moreover, sequence analysis revealed an extra guanosine nucleotide, one of the subunits of RNA, at the beginning (what is called the 5′, or upstream, end) of the linear molecule. It is extra in the sense that it is not encoded in the DNA from which the RNA is transcribed.

Strangest of all, when Cech and his co-workers tried to follow in detail the transcription from DNA to RNA, they found the precursor RNA was short-lived. Even in a laboratory system that should not have promoted splicing, the intervening sequence was quickly excised and circularized. No protein seemed to be involved in the process. By manipulating the concentration of salts in their preparation the investigators were eventually able to separate transcription and splicing, so that unspliced precursor RNA accumulated. When they added a supply of guanosine triphosphate (GTP), excision and circularization followed, even when proteins were carefully excluded from the preparation. The supplied GTP was shown to be mediating the excision by inserting itself at the junction between the upstream exon and the beginning of the intervening sequence. Another guanosine nucleotide at the downstream end of the intervening sequence then inserts itself near the 5′ end to circularize the excised molecule.

The Colorado group has now reported more rigorous evidence of autoexcision and autocircularization. Writing in *Cell,* Kelly Kruger, Paula J. Grabowski, Arthur J. Zaug, Julie Sands, Daniel E. Gottschling and Cech report on experi-

Pocket Big Savings On This Radio Shack TRS-80® Hand-Held Computer System

TRS-80 Pocket Computer Model PC-2

Cut $80 **199.95**

Was $279.95 in Cat. RSC-8

New Low Price! Now the powerful Radio Shack PC-2 (Cat. No. 26-3601) is more affordable than ever! Engineers, students, scientists—anyone with tough problems to solve—can put this TRS-80 computing system to work.

Unexpected Power. The PC-2's RAM is 2640 bytes big—expandable to over 10K. Our phenomenal Extended Pocket BASIC programming language gives you comprehensive handling of string variables up to 80 characters long, 2-dimension arrays, and 10 digit accuracy. It includes 42 statements, 34 functions, and 6 commands.

Superior Speed. The PC-2's 8-bit CPU operates at a clock speed of 1.3MHz—as fast as many desktop computers. Its advanced design allows direct memory access, and maskable and timer interrupts.

Additional Features. PC-2 has a built-in real-time quartz clock, accessible from either BASIC or the keyboard. The 26-character Liquid Crystal Display can produce full upper and lower case alphanumerics, special engineering and scientific characters or 7 × 156 dot matrix graphics. There's even a programmable beeper.

Printer/Plotter Cassette Interface

Cut $20 **219.95**

Was $239.95 in Cat. RSC-8

Flexibility With Peripherals. Add our unique Printer/Plotter/Dual Cassette Interface (26-3605)—also at a new low price—and plot full X/Y/Z axis graphics in four colors! Add two cassette recorders for automatic overlays, chaining and data storage under program control.

See It Today. Visit your nearest Radio Shack Computer Center, store, or participating dealer—where Pocket Computers start at just $149.95.

Radio Shack®

The biggest name in little computers®

A DIVISION OF TANDY CORPORATION

Send me your free TRS-80 Computer Catalog today!

Mail To: Radio Shack, Dept. 83-A-273
300 One Tandy Center, Fort Worth, Texas 76102

NAME _____

ADDRESS _____

CITY _____ STATE _____ ZIP _____

PHONE _____

Retail prices may vary at individual stores and dealers.

ments in which the DNA of the intervening sequence and parts of its adjacent exons were inserted into a plasmid (a small circle of bacterial DNA). The purified plasmid DNA was transcribed with a bacterial polymerase, the enzyme that assembles RNA nucleotides on a DNA template. Then the polymerase was removed, so that no protein remained in the purified preparation. Again the intervening sequence excised itself and formed itself into a circle. Cech and his colleagues were also able to show that the two exons are ligated in the course of the same reaction that excises the intervening sequence. Several lines of evidence suggest that in many respects the *T. thermophila* intervening sequence functions as an enzyme. Cech and his colleagues call it a ribozyme.

How prevalent are ribozymes? No one knows. Few splicing enzymes have been identified and it could turn out that autocatalysis is widespread (although it does not seem likely to be responsible for the splicing of messenger-RNA precursors, which seems not to depend on the integrity of the intervening sequences). The discovery of just one instance of self-splicing is enough to stimulate new speculation about RNA, however. Unlike the double helix of DNA, a single strand of RNA can fold readily into a variety of shapes, and a typical enzyme is characterized primarily by a highly specific, complex shape. For several years the suspicion has been growing that some RNA's may have an active rather than a purely structural role in translation and some other cellular reactions. In addition to transfer RNA and ribosomal RNA there are many small RNA-protein complexes in cells, some of whose functions are just beginning to be glimpsed. Many people who think about the origin of life favor the hypothesis that RNA was the first genetic material, antedating DNA. The ability of primitive RNA's to catalyze their own rearrangement would presumably have given the early genetic material a large capacity for variation, the primary requirement for evolution.

Toward the Silicon Foundry

According to Gordon E. Moore of the Intel Corporation, the microelectronics industry obeys two laws of growth: the number of circuit elements that can be fitted on a chip of silicon doubles each year, and the cost of designing such a chip increases almost as fast as its complexity. A large-scale integrated circuit such as a microprocessor can have more than 450,000 circuit elements and can take two years or more to develop. The cost of designing the device can exceed that of fabricating all the chips made from the design.

In response to rising costs new and more structured approaches to integrated-circuit design are being developed. In addition computer-aided design systems are being introduced to relieve the designer of the need to describe an integrated circuit at its most primitive and detailed level. The result may be a major change in the semiconductor industry. System designers, who now build electronic equipment from the chips designed by commercial semiconductor manufacturers, may become designers of their own integrated circuits. And a proliferation of custom chips, tailormade for specific low-volume applications, may to some extent replace general-purpose, high-volume chips.

At present electronic systems are customized largely by programming. A typical product consists of a few large-scale, programmable chips, such as microprocessors and memories, and many simpler chips needed to support the operation of the more complex ones; all the parts are assembled and interconnected on a printed-circuit board. The builder of the system adapts it to a specific purpose by choosing the parts and the interconnection scheme, but the system's function is determined primarily by programming the microprocessors and memories. What the system designer has not been able to do is alter the internal structure of a chip.

In the past five years a variety of semicustom technologies have begun to enable the system builder to intervene in the design of some types of integrated circuit. One semicustom device is the gate array, a chip on which thousands of logic gates (combinations of a few circuit elements that perform a simple logical function) are arranged in regular rows. The gates themselves are the same on all the chips, but the layers of metallic conductors that determine which gates are used and how they are connected can vary from one chip to another. The system designer defines the function of the chip by specifying the pattern of connections.

About 40 companies are now making gate arrays. Some of them provide computer-aided design systems that help to route the connections in a way that exploits the resources of the chip efficiently. A typical design system includes a library of the connection patterns corresponding to elementary units of electronic logic. The designer enters a description of the logical functions the device is to carry out and the system translates the description into a pattern of connections. The pattern serves as a template for the deposition of the metallic layers.

The building blocks of a gate array are logic elements. The building blocks of another semicustom design technology, called the standard cell, include logic elements, memories and central processing units. Some eight semiconductor manufacturers offer standard cells. The manufacturer defines the functions of a cell and creates a data base giving its characteristics, such as the time needed for a signal to propagate through it. Computer-aided systems assist the designer in combining various standard cells on a single chip. The designer breaks the system down into blocks corresponding to cells and creates network lists that specify the cells and how they are to be connected. The computer-aided design system positions the cells and routes the connections between them. Ultimately the system generates the data from which all the layers of the chip will be made.

Standard cells allow more intervention in the design of a chip than gate arrays do, but the system designer still cannot control the internal layout of the cells themselves. Carver A. Mead of the California Institute of Technology and Lynn A. Conway of the Xerox Palo Alto Research Center have created a structured approach to the custom design of integrated circuits that completely frees the system builder from dependence on the design expertise of semiconductor manufacturers. An automated design system supporting this approach would generate the entire layout of each chip rather than retrieving blocks of the layout from a data base.

Mead calls an automated design system of this kind a silicon compiler by analogy to the compiler program that supports a high-level computer language. The silicon compiler accepts a high-level functional definition of a device and translates it automatically into a chip layout in a standard format, much as a computer-language compiler translates programs written in a high-level language into instructions that can be executed by a computer. Several universities have developed silicon compilers and the first commercial one is to be introduced soon.

Silicon compilers could make it possible for a large population of engineers to design their own integrated circuits. As a result, Mead notes, a new kind of semiconductor manufacturer could begin to flourish: the silicon foundry. A silicon foundry would handle only the fabrication of a chip, leaving the design to its customers, much as a metal foundry casts parts designed by customers.

By processing many low-volume chip designs a silicon foundry would be able to compete with a commercial manufacturer designing and processing a few high-volume chips. Some "foundry" business is now done by commercial manufacturers, which interleave production of customer-designed chips with production of their own designs. True silicon foundries are also beginning to appear. As both silicon compilers and silicon foundries make custom design more economical, electronic systems for a variety of applications may

be systems on a single chip rather than systems assembled from many chips on a printed-circuit board.

Detumescence

When earth movements of any kind are reported in fault-checkered southern California, everyone listens. The citizens of Palmdale, a small agricultural community astride the San Andreas fault zone 35 miles northeast of Los Angeles, have been particularly attentive. In 1976 three geologists reported in *Science* that the town was nearly a foot higher than it had been in 1955. One of the authors of the report, Robert O. Castle of the U.S. Geological Survey, later revised his estimate of the degree of uplift to more than a foot. Half of the uplift, Castle noted, seemed to have taken place between 1961 and 1965. The movement could not be related to any local seismic activity, but the possibility was considered that it might be a sign of an impending earthquake.

Speculation about the cause of the Palmdale "bulge" has continued since then, but no convincing explanation has been given. Now Sandford R. Holdahl of the National Oceanic and Atmospheric Administration has suggested that no explanation is needed because in fact there has been no significant uplift. To account for the earlier findings he cites a bit of surveyors' wisdom propounded in 1981 by a fellow geophysicist, W. E. Strange: Do not conduct leveling surveys along railroad tracks in the southwestern U.S., where solar radiation levels are high.

Writing in *Journal of Geophysical Research,* Holdahl reconstructs the story as follows. Until 1964 the instrument used by surveyors in California for measuring elevation was good for sighting over long distances. It was customary for surveyors to make their measurements along railroad tracks, where the gradual slope allowed such long sightings. Surveyors even traveled from sighting to sighting on handcars. It had been known since the 1930's that the leveling procedure is susceptible to refraction errors: a temperature gradient in the air along the sight line can bend the path of light and give an inaccurate measure of elevation. Nevertheless, surveys in the U.S. were not corrected for refraction because the magnitude of the errors was thought to be small.

In most instances the errors probably are small. Both railroad ballast and the rails themselves, however, are efficient heat sinks; air-temperature gradients in their vicinity are large, particularly in the sunny Southwest. Surveys along tracks were thus exposed to significant refraction error, and the problem was compounded by the long sightings.

The introduction of new level instruments in 1964–65 resulted in a reduction of sighting distances and so of refraction error, even when surveys were done along rail rights-of-way. Holdahl recalculated the leveling measurements that had led to the proposal of a bulge at Palmdale, correcting for refraction error. The rise proved to be not a foot but somewhat less than three inches (plus or minus 1.6 inches). Even the slight remaining apparent uplift might be due at least in part to residual error.

Topological Perception

Current attempts to understand vision often begin by affirming that the brain analyzes images for their content of rather abstract features. A notable example is the detection of edges by nerve cells in the visual cortex. Another example may be the detection of topological properties such as connectedness. Writing in *Science,* Lin Chen, who holds positions at the China University of Science and Technology in Hofei and at the University of California at San Diego, suggests that the detection of topological properties may turn out to be "a primitive and general function of the visual system."

Chen reasoned from experiments showing that people readily distinguish a figure from its background even when "the detailed structure of the stimulus remains vague and amorphous." The distinction must have depended only on the overall properties of the stimulus. Such properties, Chen notes, "can be described mathematically as topological properties, such as connectivity." In topology a circle, a triangle and a square are equivalent. Any two points in such a figure can be connected by a straight line that stays entirely inside the figure. On the other hand, a circle, a triangle and a square are quite different from a ring or any other figure that has a hole in it.

In one experiment Chen had subjects press a button in order to see a stimulus consisting of a circle and a square, a circle and a triangle or a circle and a ring. Each stimulus appeared for only five milliseconds; then the subject was asked whether the two figures in it were the same or different. Under these circumstances the visual system proved to be most sensitive to the difference between the circle and the ring; they were "reported as 'different' on 64.5 percent of the presentations." The circle and the square were "different" on 43.5 percent and the circle and the triangle on only 38.5 percent.

Another experiment presented subjects with two glimpses of a short line segment. The position and the slope of the segment were changed from the first glimpse to the second; then the subjects were asked which segment was first in the sequence. They were correct on 59.2 percent of the trials. In some additional trials the stimuli included a stationary circle placed so that one line segment would be inside it whereas the other segment would be outside. Since the additional circle stayed "in the same position, this context gave no clue about which of the two target lines was presented." Nevertheless, the subjects were correct about the sequence of segments in 79.1 percent of the trials. Evidently a topological property (the placement of the segment inside or outside a closed curve) had facilitated the identification.

"We do not know," Chen observes, "how this resonance to [topology] is generated." Yet questions of that nature lie "at the core of issues in perception."

A Bigger Pi

Pi, the ratio of a circle's circumference to its diameter, can be expressed approximately as 3.14 and more precisely to an indefinite number of decimal places. By the end of the 16th century pi was known to 30 decimal places. Since then many people have extended the calculation. For almost a century the record was held by the British mathematician William Shanks, who over a period of 20 years laboriously calculated by hand until he had 707 decimal places of pi; his results were published in *Proceedings of the Royal Society* in 1873–74. (In 1945 it was found that Shanks had made a mistake at the 527th place.) With the advent of the electronic computer the number of known decimal places rose rapidly, to 100,000 in 1962 and to a million in 1973. Now within a few weeks two workers in Japan have calculated pi to 2^{21} (2,097,152), to 2^{22} (4,194,304) and to 2^{23} (8,388,608) decimal places.

The Japanese workers are Yoshiaki Tamura of the International Latitude Observatory at Mizusawa and Yasumasa Kanada of the University of Tokyo. Each employed a fast digital computer and an algorithm devised some years ago by Eugene Salamin of Stanford, Calif. Tamura calculated pi to more than two million decimal places in 7.2 hours of computer time. Kanada and Tamura together, using a faster computer, calculated it to more than four million decimal places in 2.9 hours of computer time. They went on to make the calculation to 2^{23} places, an effort that took 6.8 hours on the faster computer, a HITAC M-28OH. Kanada and Tamura plan to continue the calculation to 2^{24} (16,777,216) places by the middle of this year.

The chief spur to such calculations is the search for departures from randomness in pi. It is widely believed the sequence of digits in the decimal expression of pi has no pattern, but the conjecture has not been proved. Even if pi is random overall, small-scale patterns can be expected. For example, does the sequence 123456789 appear in pi? If it does, where?

How we keep cars from becoming a meal for the scrapyard.

At 3M we make a small army of products to assemble, fix, refinish and protect cars and help them live longer, healthier lives.

We make everything from structural adhesives that firmly anchor windows in place to abrasives that are specially designed to speed body repairs.

Also there's Scotchgard Protector for carpets and seat cushions inside. And Scotchcal Films to decorate autos, trucks and other vehicles outside.

In fact, by listening to people's needs, we've pioneered over 900 products to serve the needs of the transportation industry. We developed everything from electrical connectors to composite materials that form helicopter rotor blades.

All because at 3M, we're in the business of hearing. So let us hear from you.

3M hears you...

For your free 3M Transportation Brochure, write: Department 020702/ 3M, P.O. Box 4039, St. Paul, MN 55104.

Name_____

Address_____

City_____ State____ Zip_____

Or call toll-free: **1-800-323-1718,** Operator 362. (Illinois residents call 1-800-942-8881.)

3M

The Optical Computer

A computer based on beams of light rather than electric currents might be capable of a trillion operations per second. The crucial component, an optical analogue of the transistor, has been built

by Eitan Abraham, Colin T. Seaton and S. Desmond Smith

The development of the digital computer in the past 40 years is so closely associated with the development of electronic technology that people tend to think of the computer as an inherently electronic device. Actually the operations carried out by a computer are logical and arithmetic ones, and they could be done by any of several means. Since the mid-1970's it has become apparent that the potential exists for constructing a computing device in which signals are transmitted by beams of laser radiation rather than electric currents. There is a powerful incentive for developing such an optical computer: it might operate 1,000 times faster than an electronic computer.

The fundamental components of any digital computer are switches capable of two different states of transmission. The speed of the computer in its calculations is limited by the time required for a switch to change states, among other factors. In an electronic computer virtually all the switches are transistors, and even the fastest transistors now in use cannot be made to change states in less than about a nanosecond, or a billionth of a second. An optical device analogous to the transistor could switch from one transmission state to the other in about a picosecond, or a thousandth of a billionth of a second.

We have developed an experimental version of the optical transistor that is switched by a small change in the intensity of an incident beam of laser radiation. The optical transistor, which we call a transphasor, is based on a property of certain crystals. An increase in the intensity of light causes a change in the crystal's refractive index, a measure of the extent to which light is slowed as it passes through the material. With the right combination of crystalline material and laser wavelength it is possible to utilize the change in refractive index to make a device in which the intensity of the transmitted radiation changes greatly with a very small change in the intensity of the incident beam. With our experimental devices we have already achieved switching times that appear to be approximately a few picoseconds.

The optical transistor could be employed to build computers that would process information in much the same way as electronic machines do but faster. In the more distant future the capabilities of the optical transistor could transform the organization of the computer itself. Each optical transistor could be the site of many simultaneous switching operations carried out on parallel laser beams; an electronic device operates on one signal at a time. Moreover, the crystals we work with can switch to successively higher levels of transmitted power with successive increases in the incident beam; in contrast, an electronic transistor of the kind employed in computers has only two output states. The adoption of devices with more than two stable states might lead to a new system of computer logic.

Constructing an optical computer would require a variety of circuit elements in addition to the optical transistor. In our laboratory at Heriot-Watt University we have demonstrated in experimental form many of the elements needed to make an integrated optical circuit. Formidable technical difficulties must be overcome before such a circuit could be manufactured. Nevertheless, the optical computer is an intriguing prospect for the relatively near future.

The Functions of a Switch

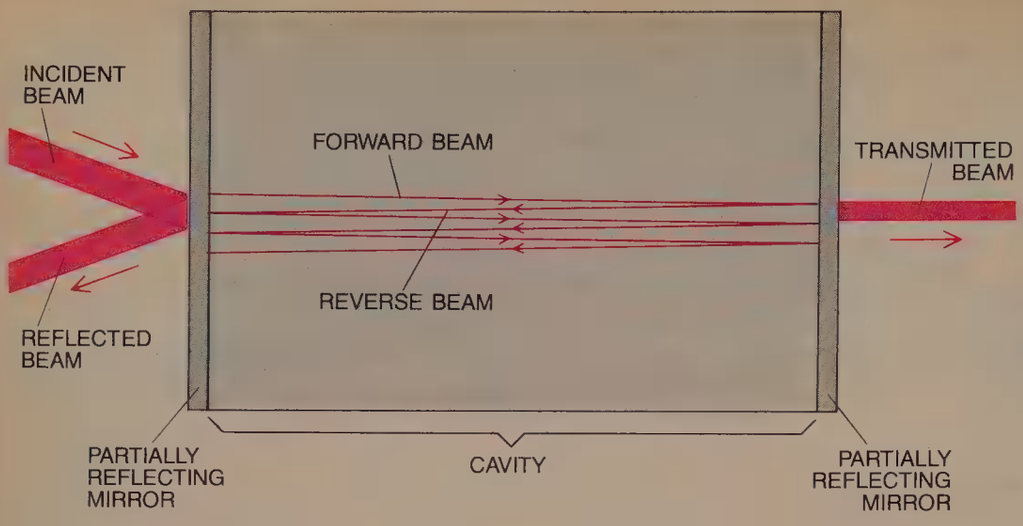
In current practice the three basic functions of a computer—arithmetic operations, logical operations and the storage of information, or memory—are all done by devices that have two stable states. In arithmetic operations the two states represent the numerals 0 and 1 of the binary number system. In the evaluation of logical propositions the two states stand for true and false. The memory of the computer stores the results of arithmetic and logical operations in devices that occupy one of two states.

With the binary algebraic system a computer can evaluate the truth of propositions by making use of just three logical functions, which are usually referred to as the AND function, the OR function and the NOT function. In the AND function a statement is taken to be true if all its components are true. In the OR function a statement is taken to be true if any of its components is true. In the NOT function the truth value of a statement is reversed. More elaborate logical operations can be built up out of the three basic functions, and so can arithmetic operations such as addition.

Thus a computer requires a device that can represent the values 0 and 1 or true and false in physical form and can be assembled into larger-scale devices that perform the three logical functions. The device obviously must have two output states that can be readily distinguished. If the computer is to operate at high speed, the time needed for switching between the output states must be short. Other considerations suggest the most valuable switch is the one that is smallest, simplest to fabricate and requires the least power.

The first practical electronic switch was the vacuum tube. It was slow and large, and its operation generated a substantial amount of heat, which limited the operating life of the tube. As a result computers based on vacuum tubes were large in physical dimensions but small in computing capacity. More powerful machines would have required continuous replacement of defective tubes.

The invention of the transistor in 1947 provided a smaller, faster and more efficient switch. In its simplest form the transistor is a three-layered sandwich of semiconductor materials. (A semiconductor is a material whose electrical conductivity is between that of a conductor and that of an insulator; the electrical properties of a semiconductor are readily amenable to alteration.) The outer layers are called the emitter and the collector; the middle layer is the base. The electrical properties of the layers are chosen so that a slight change in the current flowing from the base to the collector causes a much larger change



FABRY-PEROT INTERFEROMETER, shown in schematic form, is the basis of optical switches constructed by the authors. The device consists of two plane mirrors placed parallel to each other with a transparent material inserted into the space between them. The space is called the cavity. Each mirror reflects part of the light that strikes it and transmits the rest. Thus when a beam is trained on the front mirror, some of the light is reflected and some passes into the cavity. The transmitted "forward" beam travels through the cavity and strikes the rear mirror, where some light passes out of the device and some is reflected as a "reverse" beam. The beam continues to rebound until all the light leaks out of the cavity. Changing the length of the cavity, the material in it or the wavelength of the light can cause substantial changes in the transmission of the interferometer. In an optical switch the medium in the cavity has properties that allow the transmission to be controlled by the intensity of the incident beam.

in the current from the emitter to the collector.

A large current from the collector can stand for a 1 and a small current can stand for a 0. By combining transistors and other circuit elements, structures that carry out the AND, OR and NOT functions can be assembled; the circuits are called logic gates. The AND gate yields a large current if all its inputs are large. The OR gate yields a large current if any of its inputs is large. The NOT gate yields a large current with a small input and a small current with a large input.

Limitations of the Transistor

Depending on their size, material, design and switching power, semiconductor logic gates require from about a nanosecond to 1,000 nanoseconds to operate. Ideally, then, the fastest gates could do as many as a billion logical operations per second; in practice computer processing speeds are considerably lower. Fundamental physical considerations suggest that the present minimum of about a nanosecond is close to the shortest switching time attainable with a semiconductor switching device.

For a transistor to switch, a current of electrons must traverse the base. There are well-defined limits to the speed with which electrons can move in a semiconductor. Reducing the width of the base can reduce the switching time, but there are also limits to how narrow the base can be made. The limits have probably been reached. Therefore any substantial reductions in switching time will result not from small improvements in transistor design but from the adoption of a new switching technology.

One possibility is to replace the current of electrons with a signal of another kind to carry information through the circuit. The highest velocity a signal that carries information can have is the speed of light, and so light (or electromagnetic radiation in another part of the spectrum) is a suitable candidate for a new information carrier. Following this line of reasoning we have utilized the coherent radiation from a laser to operate switching devices with many of the capabilities of the transistor. The devices can switch between two readily distinguishable output states; they can serve as memory elements, and by a suitable choice of materials and laser beams fast, simple logic gates can be assembled.

The starting point of the optical transistor is an ingenious and widely used piece of optical apparatus known as the Fabry-Perot interferometer, invented by the French physicists Charles Fabry and Alfred Perot in 1896. Fabry and Perot employed the interferometer to measure the wavelength of various colors of light; it has since been put to many other uses. The Fabry-Perot apparatus in its simplest form consists of two plane mirrors placed parallel to each other and separated by a space. A material that transmits radiation with the wavelength of interest can be inserted into the space, which is called the cavity.

Each of the mirrors partially reflects and partially transmits the light that falls on it. Partially reflecting mirrors are commonplace. Indeed, a plate-glass display window can act as such a mirror: the reflected light provides an image of the observer and the transmitted light provides an image of the goods behind the window. The strength of the images depends on the proportion of the incident light that is reflected and the proportion transmitted.

Ignoring for the moment the material in the cavity, consider what happens when a beam of light strikes the mirror that forms the front face of the interferometer. Assume that the mirrors reflect 90 percent of the incident light and transmit 10 percent. (Such proportions are quite close to those in some of our work.) Therefore when the beam of light strikes the front mirror, 90 percent of it is reflected and 10 percent passes into the interior of the interferometer.

The light, now with a tenth of its original intensity, travels to the rear mirror as the "forward" beam. The properties of the rear mirror are the same as those of the front mirror. Thus at the rear mirror 90 percent of the forward beam is reflected back into the cavity as the "reverse" beam and 10 percent is transmitted from the interferometer. Because the forward beam is a tenth as strong as the original incident beam the transmitted light has an intensity a hundredth as great as that of the original beam. The reverse beam, of course, travels back to the inner surface of the front mirror, where it is divided again into a transmitted part and a reflected part. The attenuated beam continues to rebound within the cavity, becoming progressively weaker, until all the incident light leaks out of the interferometer. Each passage makes a contribution to the transmitted beam, but the contributions after the first one are quite small; as an approximation it is reasonable to think of the transmitted light as being about a hundredth of the incident beam.

Interference in the Cavity

If the account given above captured all that goes on in the interferometer, the device would be of little interest as a switch for an optical computer. The intensity of the transmitted beam could be increased only by altering the properties of the mirrors. There is, however, an additional consideration. The forward and reverse beams cannot be separated in reality as neatly as we have separated them in our schematic description. In an actual interferometer the interaction of the beams influences the intensity of the light in the cavity, which in turn determines the transmission.

As is well known, light has some of the characteristics of a wave and some of those of a particle. Sometimes one description is more convenient and sometimes the other; in this context the wave description is more appropriate. The forward beam and the reverse beam along with the successive weaker rebounds can be thought of as waves propagating through the cavity. The incident beam enters the interferometer perpen-

dicular to the front face; therefore all the beams travel along the same path.

Keeping in mind that there are many reflected beams in the cavity, we can consider the interaction of just two of them: the forward beam and the reverse beam. The outcome of the interaction depends on the phase of the waves, that is, on the relative positions of the crests and the troughs. Crests can be aligned with crests and troughs with troughs in the relation known as constructive interference. Alternatively, the crests of one wave can be aligned with the troughs of the other in destructive interference. Of course, every phase between these extremes is also possible.

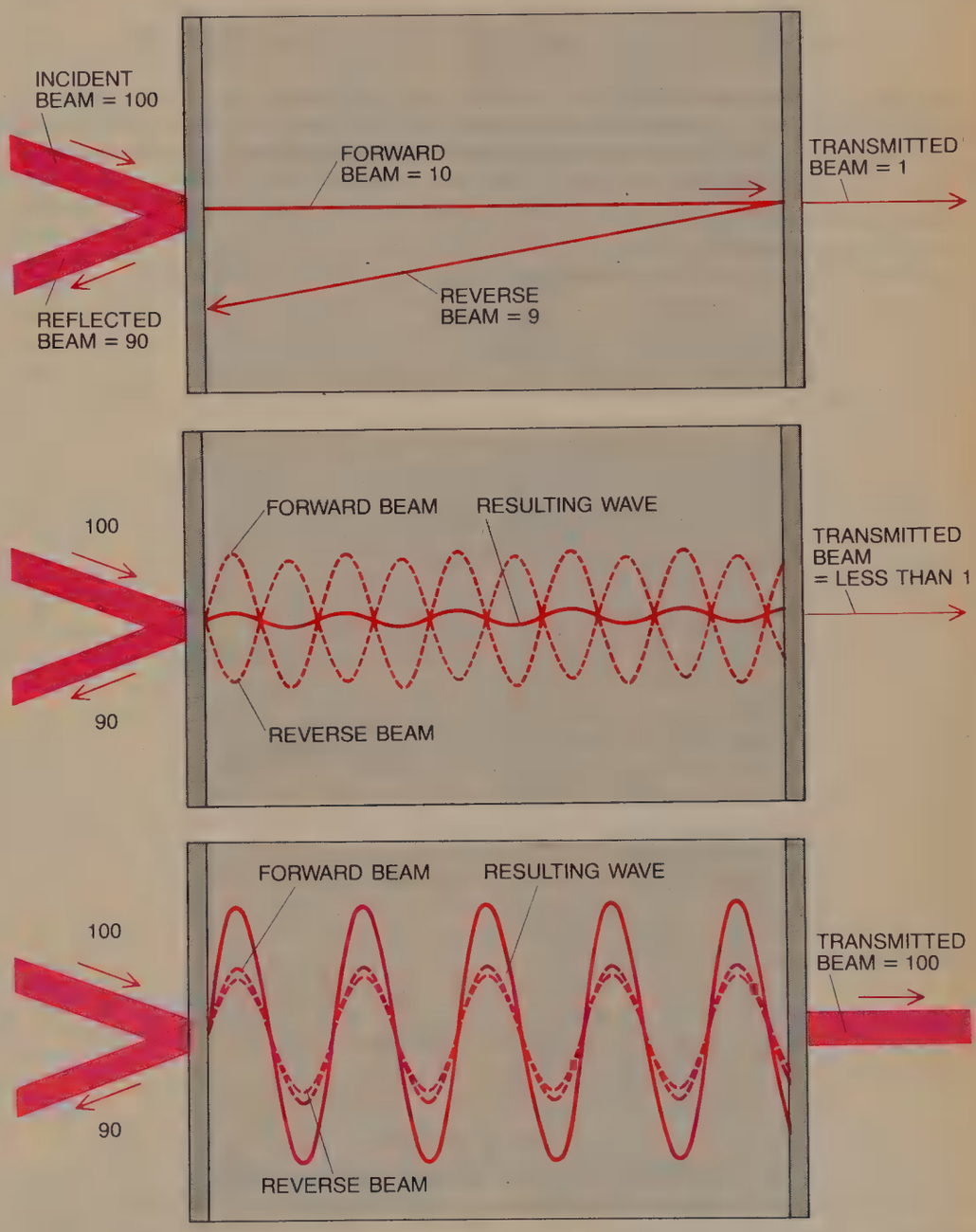
Through interference the two overlapping waves combine to form a third wave. The amplitude of the resulting wave at each point in the cavity is the sum of the amplitudes of the components at that point. If the interference is constructive, the amplitude is large; if the interference is destructive, the amplitude is small. The effect of interference on the transmission of the Fabry-Perot interferometer is far from trivial. At full destructive interference the intensity of the light in the cavity is almost zero and transmission is negligible. At full constructive interference the many forward and reverse beams reinforce one another and yield a light intensity in the cavity that is as much as 10 times the intensity of the incident beam. Since the rear mirror transmits 10 percent of the light reaching it, at full constructive interference the interferometer transmits a beam equal to the incident beam.

Effect of the Refractive Index

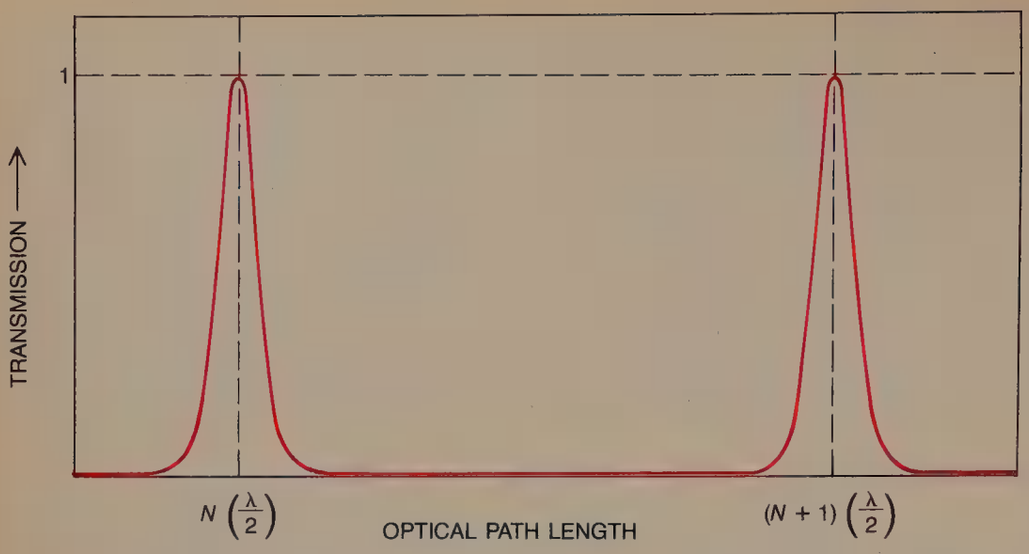
The phase of the forward and reverse beams, and hence the transmission of the interferometer, can be adjusted by changing the length of the cavity or the wavelength of the incident light. Neither method is of much value for the optical transistor. What is more important, the phase relations can also be changed by altering the optical properties of the material in the cavity, which up to this point we have passed over.

Light attains its maximum speed of propagation in a vacuum. In a transparent medium other than a vacuum the light is slowed by an amount that varies with the medium. It is the slowing of the light wave that gives rise to the familiar phenomenon of refraction; the refractive index of a material expresses the ratio of the speed of light in a vacuum to its speed in the material. For example, if the refractive index of a certain kind of glass is 1.5, then whereas light travels at about 300,000 kilometers per second in a vacuum, its speed is about 200,000 kilometers per second in the glass. The slowing of the light is accompanied by a reduction in the wavelength, but the frequency does not change. It can readily be seen that even without a change in the length of the cavity or the wavelength of the incident light a change in the refractive index of the material in the cavity will alter the phase relations of the beams in it by altering the wavelength. By selecting a material with the right refractive index the forward and reverse beams can be made to interfere either constructively or destructively.

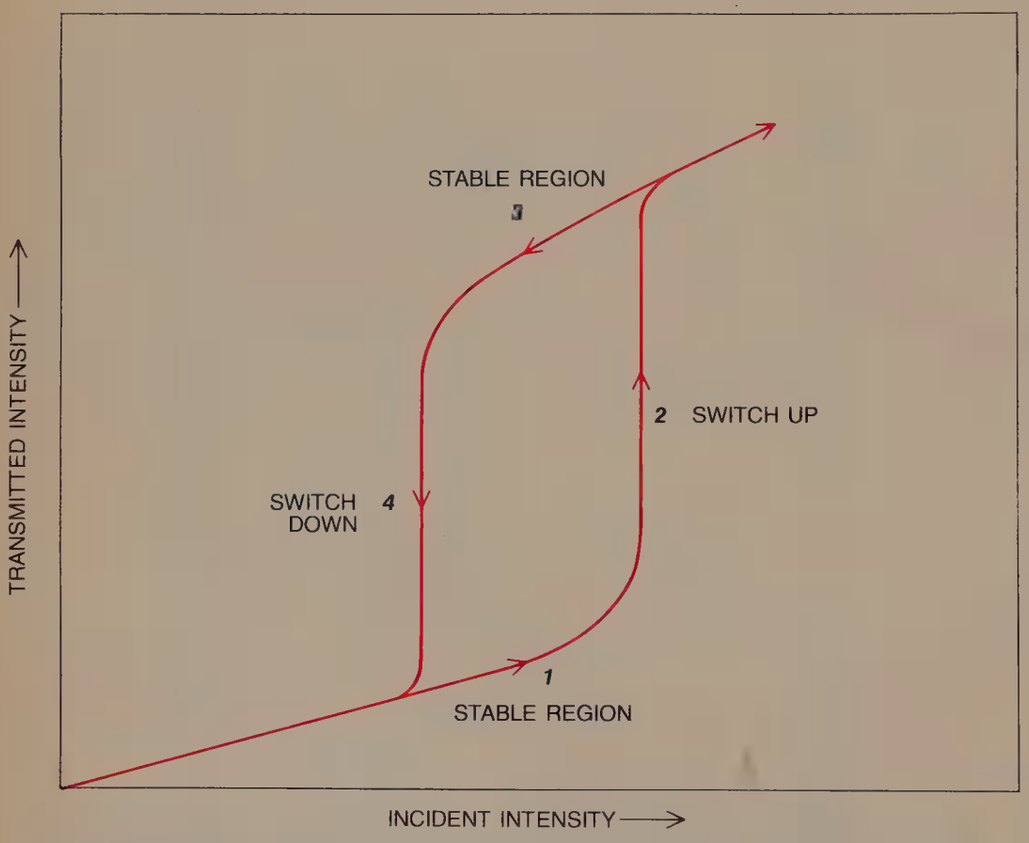
An important quantity in analyzing an optical system is what is termed the optical path length, or simply the optical length. The optical length of the medium in the interferometer cavity is its physical length multiplied by its refractive index. It can be shown that full constructive interference results when the optical length is equal to the distance represented by an integral number of half wavelengths of the incident light. (A half wavelength, of course, is the distance from a crest to a trough.) There are thus many possible optical lengths that yield full constructive interference. When the optical length is halfway between distances corresponding to integral multiples of the half wavelength,



INTERFERENCE of the forward and reverse beams in the Fabry-Perot cavity can greatly change the light intensity within it and hence the transmission of the device. The front and rear mirrors can be chosen so that they reflect 90 percent of the light striking them and transmit 10 percent. Without considering interference the situation is as follows (*top*). Ten percent of the incident beam passes into the cavity. When the attenuated beam strikes the rear mirror, a beam with 1 percent of the original intensity is transmitted; the reverse beam is equal to 9 percent of the original one. Actually the forward and reverse beams interact as waves to yield a wave equal at each point to the sum of the two components. If the crests of one beam are aligned with the troughs of the other in destructive interference, the resulting wave is small (*middle*). Light intensity in the cavity is negligible and transmission can be less than 1 percent of the incident beam. If the two waves are aligned in constructive interference, the resulting wave is large (*bottom*). Because there are many forward and reverse beams, the intensity in the cavity can be as great as 10 times that of the incident beam. Thus when the interference is fully constructive, the transmitted beam can be equal in intensity to the original incident beam.



AIRY FUNCTION expresses the relation between the transmission of the interferometer and the optical path length of the material in the cavity. The transmission is the ratio of the intensity of the incident beam to that of the transmitted beam; its maximum is 1. Transmission is at the maximum when there is full constructive interference between the beams in the cavity. The optical path length is the physical length of the material in the cavity multiplied by its refractive index. (The refractive index is a measure of the extent to which light is slowed as it passes through a transparent substance.) The interference is fully constructive when the optical length is equal to any integer multiplied by half the wavelength (λ) of the incident light.



HYSTERESIS CYCLE represents the way the intensity of the beam transmitted by the interferometer changes in response to changes in the intensity of the incident beam. In the cavity is a material with a nonlinear refractive index: one that varies according to the intensity of the light in the material. Changes in the incident beam alter the refractive index and the optical length and hence the transmission. If the intensity of the incident beam is increased slowly, the transmitted intensity initially shows a gradual increase (*1*). The change in refractive index, however, soon brings the optical path length near the value corresponding to constructive interference, thereby increasing the light intensity in the cavity; the change in intensity further alters the refractive index. At a certain point this positive feedback becomes quite strong. A threshold is reached and transmission increases rapidly (*2*). If the incident intensity is then reduced somewhat, transmission does not fall to the original level because there is still enough light in the cavity to keep the refractive index and the optical path length near the values corresponding to maximum transmission; transmission falls gradually (*3*). At a second threshold, however, further dimming of the incident beam changes the refractive index enough to cause a greater loss of intensity in the cavity; transmission then falls rapidly to the initial low level (*4*). An optical device that has a hysteresis cycle is said to be optically bistable. Regions *1* and *3* of the curve are the stable regions, where transmission changes little with changes in intensity.

the result is full destructive interference.

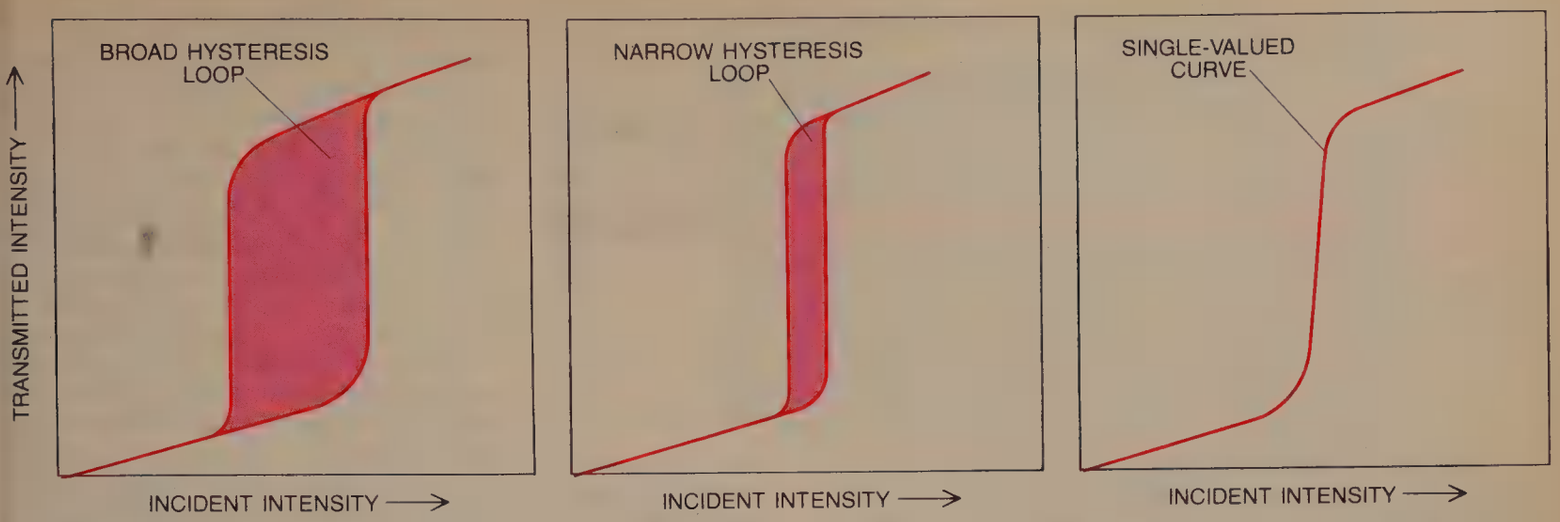
Since the intensity of the light transmitted by the interferometer varies according to the phase relations of the beams, it follows that the transmitted intensity varies with the optical length. The function that describes the relation between the optical length and the transmission of the interferometer (defined as the ratio of the strength of the transmitted beam to that of the incident beam) is called an Airy function and is quite important in the construction of an optical transistor. In the Fabry-Perot devices we have been building the function has a steep peak at each point where the optical length corresponds to an integral number of half wavelengths; between the peaks the value of the function is small, and it changes only gradually. Therefore as the optical length approaches a value corresponding to peak transmission, the transmission increases slowly until a threshold is reached and then increases very rapidly.

The effect of the refractive index on the optical length implies that the transmission of a cavity can be set at some fixed level simply by inserting into the cavity a material with the appropriate refractive index. Moreover, exchanging one material for another would alter the transmission. Of course, such a method of switching would be quite clumsy in a computer. In the 19th century, however, it would have been considered the only means of altering the refractive index. James Clerk Maxwell, in his fundamental work on electromagnetic radiation, postulated that refraction and other interactions of light with matter are independent of the intensity of the light. If this were always true, the refractive index would be a constant and the output of the interferometer would be directly proportional to the intensity of the incident beam. Under the appropriate conditions many substances yield such a result; they are said to be linear with respect to the refractive index.

Nonlinear Refraction

When the development of lasers made available a powerful source of coherent radiation (that is, radiation in which all crests and troughs are aligned), it became apparent that not all materials have a linear refractive index. When laser radiation is focused on certain substances, the refraction varies with the intensity of the beam; the materials are nonlinear with respect to the refractive index. With a nonlinear substance in the Fabry-Perot cavity the refractive index can be altered merely by changing the intensity of the incident beam.

The effect is subtle enough to be considered in detail. Imagine a cavity of fixed length with a nonlinear material in it; a beam of laser radiation whose intensity can be adjusted is directed into the



HYSTERESIS LOOP CAN BE ALTERED in an optically bistable device by changing the length of the interferometer, the wavelength of the incident beam or the material in the cavity. In the basic hysteresis cycle the stable regions are wide and the loop is broad (*left*). A given level of incident intensity can correspond to either one of two levels of transmitted intensity. By the appropriate manipulation of the interferometer the stable regions can be reduced and the loop narrowed (*middle*). The limit of the process is a single-valued curve, in which each level of incident intensity corresponds to only one level of transmitted intensity (*right*). The curve is similar to the transfer-characteristic curve of an electronic transistor; an interferometer with such a transmission curve is utilized in the optical transistor.

cavity. Initially the intensity of the beam is such that the refractive index yields an optical length in the region of the Airy curve between peaks; transmission is low. If the intensity of the beam is slowly increased, the refractive index and the optical length gradually change; transmission increases slowly. At some point the refractive index and the light intensity in the cavity form a positive-feedback loop, that is, they become mutually reinforcing. The changing refractive index brings the beams closer to constructive interference, thereby augmenting the intensity in the cavity; the increased intensity further changes the refractive index. At a certain critical level of incident intensity the feedback effect becomes quite strong. The steep region of the Airy curve is reached and transmission increases very rapidly indeed.

At the peak of the Airy function the ratio of transmitted intensity to incident intensity is 1. When the intensity of the laser beam is reduced somewhat from the peak value, the transmission of the cavity does not fall to its initial low level. The reason is that the light intensity in the cavity is sufficient to keep the refractive index and the optical length near the values that yield maximum transmission. As the intensity of the incident beam is reduced further, transmission still decreases only gradually. At a certain point, however, the refractive index and the intensity in the cavity become mutually attenuating. A small decrease in the incident intensity changes the refractive index enough to reduce the intensity in the cavity substantially. Hence there is a sudden sharp decrease in transmission.

If the intensity of the transmitted radiation is plotted against the intensity of the incident beam, the resulting curve is a loop with four sections [*see bottom illustration on opposite page*]. Beginning at a low incident intensity the curve rises gradually: even a large increase in incident intensity causes only a small gain in transmission. At the threshold the curve becomes steep: a small increase in intensity brings a sharp gain in transmission. The return to the starting point follows a different path. Initially the curve slopes downward only slightly, and even if the intensity is reduced, the transmission remains high. When a second threshold is reached, however, the curve falls precipitously: a small loss of intensity greatly attenuates the transmission.

The loop constitutes what is called a hysteresis cycle. An optical device that exhibits such a cycle is said to be optically bistable because it has two stable regions where the transmitted intensity varies little with changes in the incident intensity. It should be noted that in a bistable device a particular value of incident intensity can be associated with two levels of transmitted intensity; the actual transmission depends on the previous condition of the cavity.

Bistable Optical Devices

The existence of optical bistability was predicted by Abraham Szöke and his colleagues at the Massachusetts Institute of Technology in 1969. It was first observed by Hyatt M. Gibbs, Samuel L. McCall and Thirumalai N. C. Venkatesan of Bell Laboratories in 1976. Since then bistable optical devices have been the subject of work in many laboratories.

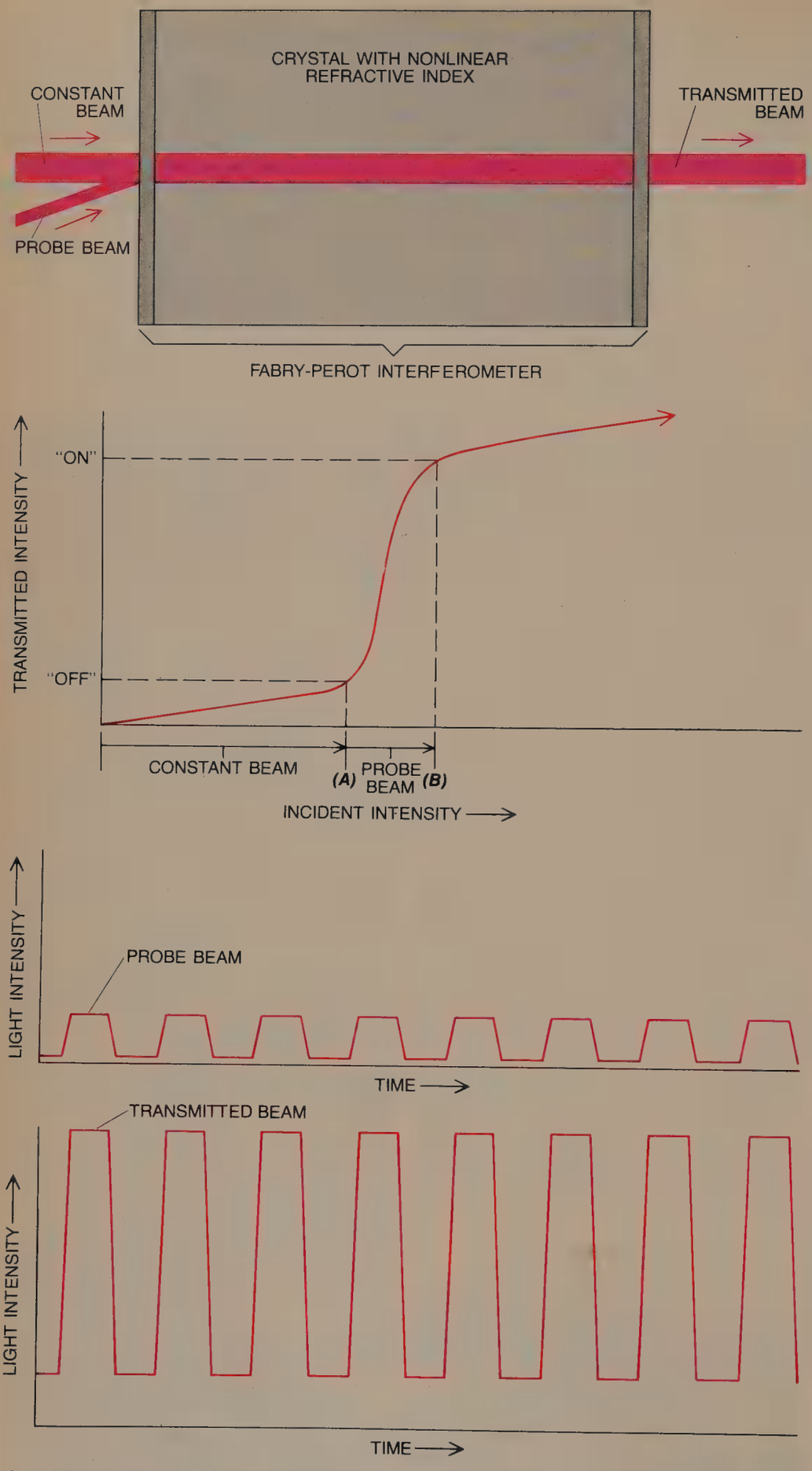
In making an optical bistable device it is important to choose a material with a substantial nonlinearity in its refractive index. In our work we have utilized the compound semiconductor indium antimonide, which has a highly nonlinear response. Like most semiconductors, indium antimonide is opaque to radiation in the visible part of the spectrum, but it is transparent to infrared radiation of some wavelengths. Our source of radiation is a carbon monoxide laser whose wavelength can be adjusted over a narrow region in the infrared.

The interferometer we employ is a small rectangular crystal of indium antimonide. Generally each edge of the crystal is a few millimeters long; we have also built devices out of much smaller crystals. In some cases the front and rear faces of the crystal are coated with a reflective substance that serves as the mirror, but the coating is not essential: the polished face of the crystal itself can serve as the mirror. A device beyond the rear mirror measures the intensity of the transmitted beam.

On the atomic scale there are many mechanisms that can change the refractive index of a semiconductor. The mechanism in indium antimonide is complex and indeed is still being worked out. Some aspects of the process are understood; the following is a simplified account of what is known.

The interaction of electromagnetic radiation with a solid material is mediated almost entirely by the electrons in the solid; at the wavelengths of visible light or infrared radiation it is mainly the outermost electrons of each atom that participate. In a good conductor of electricity, such as a metal, the outermost electrons are free to roam throughout the volume of the material; in a good insulator each electron is bound tightly to a particular atom and can be dislodged only at a high cost in energy. The situation in a semiconductor is intermediate between these extremes: each of the outermost electrons is bound to an atom but only a small input of energy is needed to set an electron free.

These differences in electronic configuration, which determine the optical as



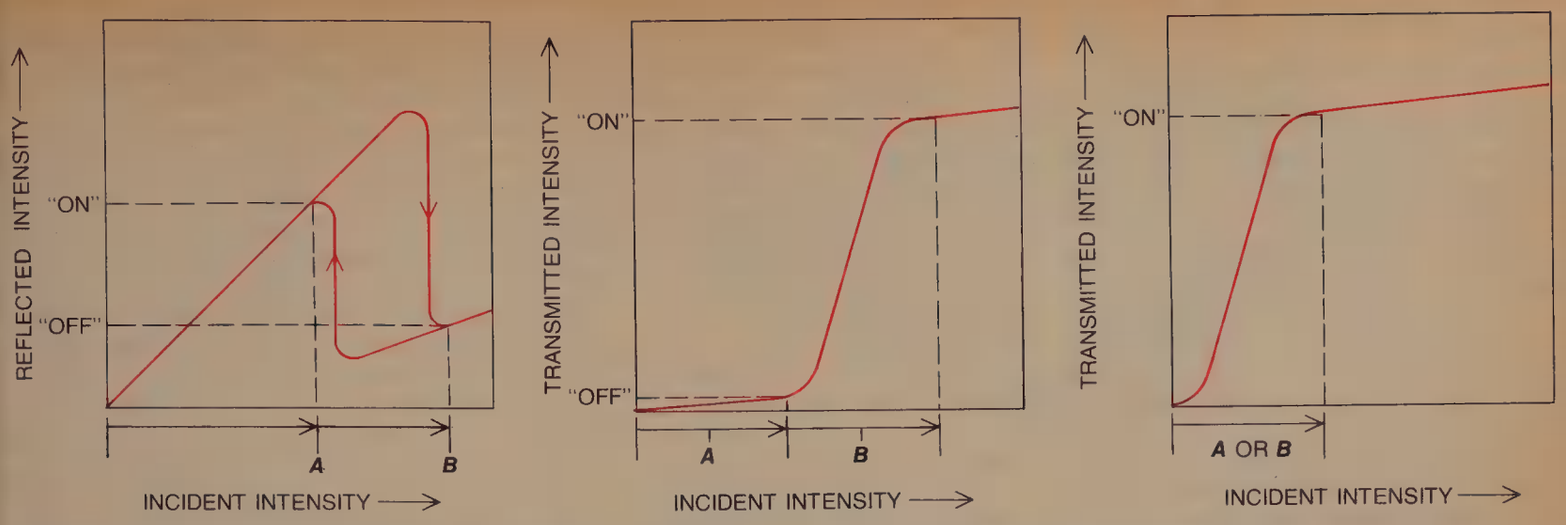
OPTICAL TRANSISTOR devised by the authors relies on two precisely adjusted laser beams focused on the front face of a Fabry-Perot interferometer containing a substance with a nonlinear refractive index (*top*). The "constant" beam is strong and unvarying. The "probe" beam is weaker and can be modulated. The intensity of the constant beam (*A*) is chosen to yield a low transmitted intensity that is just below the steep region of the single-valued transmission curve (*middle*). The intensity of the probe beam (*B*) is chosen so that when it is added to the intensity of the constant beam, the transmission increases to the high level at the top of the steep region. The low-transmission or "off" state can stand for the value 0 in the binary logic of the computer; the high-transmission or "on" state can stand for 1. Because of the steepness of the transmission curve, small changes in the probe beam can cause large changes in the transmitted intensity (*bottom*). In this respect the optical transistor is analogous to the electronic transistor.

well as the electrical properties of the material, can be understood by considering the energy levels available to each electron. In an isolated atom the levels are sharply defined, but in a solid the levels associated with the many atoms stack up, so to speak, to form broad bands of allowed energies separated by forbidden band gaps, representing energies no electron can have. If a band is completely filled, meaning that every level in it is occupied, each electron in the band is confined to the vicinity of an atom. The reason is that a little extra energy is needed to set an electron in motion, and there is no vacant level in the band to which an electron with extra energy could move. If a band is only partly occupied, an electron that acquires extra energy can readily move to a slightly higher level in the band and is thereby set free.

Semiconductor Band Structure

A metal has a partly occupied band. In an insulator the highest occupied band (the valence band) is filled exactly, and it is separated by a wide gap from the next higher band (the conduction band), which is empty. An electron could be set in motion in an insulator only if it absorbed enough energy to cross the gap, an unlikely event. A semiconductor also has a filled valence band and an empty conduction band, but the gap between them is smaller. There are several ways an electron in a semiconductor can gain the rather small amount of energy needed to go from one of the higher levels of the valence band to one of the lower levels of the conduction band. When an electron is promoted to the conduction band, it leaves behind it a "hole" in the valence band, which in some respects acts as an entity with a positive charge equal in magnitude to the negative charge of the electron. When the promoted electron returns to the conduction band, it joins a hole and the hole is eliminated; the process is called recombination.

The energy required for an electron to traverse the band gap varies with temperature. Much of our work with indium antimonide has been done at 77 degrees Kelvin. At this temperature the band-gap energy of indium antimonide is about .2 electron volt. (An electron volt, abbreviated eV, is the energy an electron gains when it is accelerated through a potential difference of one volt.) Most other semiconductors have a much wider band gap. The nonlinearity of the refractive index increases in inverse proportion to the square of the band-gap energy. Because of its narrow gap indium antimonide has a nonlinearity between 100 and 1,000 times as great as that of gallium arsenide, another semiconductor that has been utilized in bistable optical devices. Although a



OPTICAL LOGIC GATES have been constructed from optical transistors. A logic gate is a device employed in a digital computer for arithmetic operations and for evaluating the truth of propositions. Three kinds of gate are required; the principles according to which optical versions of them have been constructed are shown. The AND gate yields a high output only when all its inputs are present (*left*). If the two incident laser beams are chosen so that each one yields a transmitted intensity just below the steep part of the curve, the result is an AND gate: both incident beams must be present for high transmission. The OR gate yields a high output if any of its inputs is high (*middle*). If the incident beams are adjusted so that each one yields a transmission at the top of the steep region, the result is an OR gate: either incident beam can yield high transmission. The NOT gate has a high output when the input is low and a low output when the input is high. The optical NOT gate exploits the light reflected from the interferometer (*right*). The reflected beam is associated with the hysteresis loop shown. Because of the shape of the loop, a high input (*B*) yields a low output whereas a low input (*A*) yields a high output.

narrow band gap is desirable, if the gap is too narrow, operation of the optical switch becomes impractical.

One way for an electron to acquire the energy needed for promotion to the conduction band is by absorbing incident radiation. Here it is convenient to adopt the description of light as a stream of particles. The particle, or quantum, of light is the photon. The energy of a photon is inversely proportional to the corresponding wavelength. When a photon is absorbed by the crystal, the photon's energy can be added to that of an electron. If the photon carries the right amount of energy, the electron is promoted to the conduction band. The wavelength equivalent to .2 eV is about five micrometers, which is in the infrared part of the spectrum and within the range of our carbon monoxide laser.

Adjusting the laser to a wavelength of about five micrometers therefore leads to the promotion of many electrons into the conduction band. Once the electrons reach the lower levels of the conduction band the thermal energy of the crystal lattice causes them to scatter to nearby energy levels. The result is an irregular distribution of electrons in the lower reaches of the conduction band.

It is the promotion and subsequent scattering of the electrons that causes the refractive index of indium antimonide to change. In the quantum-mechanical account of how light interacts with matter, refraction and absorption are closely connected. Indeed, each entails the other: if a beam of radiation is absorbed, it is also refracted, and vice versa, although the two phenomena may affect radiation of different wavelengths within the beam. Preventing absorption from taking place can thus simultaneously prevent refraction from taking place and thereby alter the refractive index. How is absorption to be prevented? There is a limit to the number of electrons that can be promoted to the lower levels of the conduction band. Once the limit has been reached no additional photons with the precise energy needed to accomplish those promotions can be absorbed; the absorption of such photons is said to be saturated. The saturation of the absorption forestalls the accompanying refraction and the refractive index changes. Very soon after the laser beam is turned off the promoted electrons return to the valence band and recombine with holes. Absorption then becomes possible again and the refractive index reverts to its original value.

Optical Logic Gates

We have taken advantage of the extreme optical properties of indium antimonide that result from this mechanism to make components for an optical computer. The hysteresis cycle described above has an obvious application as a memory element. The stable high-transmission state of the interferometer can represent one logical or arithmetic value and the stable low-transmission state the other value. Either state can be maintained indefinitely by a beam of intermediate intensity.

For an optical switch to be used in processing information rather than storing it hysteresis is not needed. By changing the optical setup slightly the hysteresis loop can be narrowed or eliminated altogether. The result is a single-valued transmission curve, in which each level of incident intensity is associated with only one level of transmitted intensity. The shape of the single-valued curve can itself be manipulated. In one of the most useful shapes transmission is low and almost constant at low incident intensity, then at a threshold the transmission rises steeply, reaching a high level that again remains almost constant as the incident intensity increases further. The curve resembles the one that describes the current from the collector of an electronic transistor. Indeed, the curve is the basis of the optical transistor. We call the device the transphasor because its operation is based on controlling the phase of the light within it.

In the transphasor two precisely adjusted laser beams are focused on the same spot on the front face of a crystal of a nonlinear material such as indium antimonide. The "constant" beam has an intensity that is large and unvarying. The "probe" beam has an intensity that is much smaller and can be modulated. The strength of the constant beam is chosen so that it yields a transmitted intensity just below the steep region of the transmission curve. The intensity of the probe beam is such that when it is added to the intensity of the constant beam, the steep part of the curve is traversed and the transmission peak is reached.

Because the transmission curve of the transphasor is steep a very small modulation of the probe beam controls a very large increase in the transmitted intensity of the device. The analogy with the electronic transistor can be extended. The constant beam is analogous to the constant "bias" current that flows from the emitter to the collector of the transistor. The probe beam is analogous to the smaller current from the base to the collector. Just as a small change in the base current enables the transistor to conduct a much larger collector current, so a small change in the probe beam enables the transphasor to transmit a

larger fraction of the incident radiation.

Like the transistor, the transphasor can be switched between two readily distinguishable states. The switching is extremely fast. Although we have not been able to measure the switching time directly, the indirect evidence suggests it is a few picoseconds.

Once an optical transistor has been constructed, the assembly of the prototype of a logic gate is a straightforward procedure. Indeed, a single transphasor could serve as either an AND gate or an OR gate depending on the beams supplied to it. If the two incident beams are selected so that neither beam alone is sufficient to switch the device but together they are strong enough to do so, the result is an optical AND gate. Transmission is high only when both incident beams are present. If the incident beams are chosen so that either one is capable of switching the device, the result is an optical OR gate. The transmitted intensity is high when either beam is present.

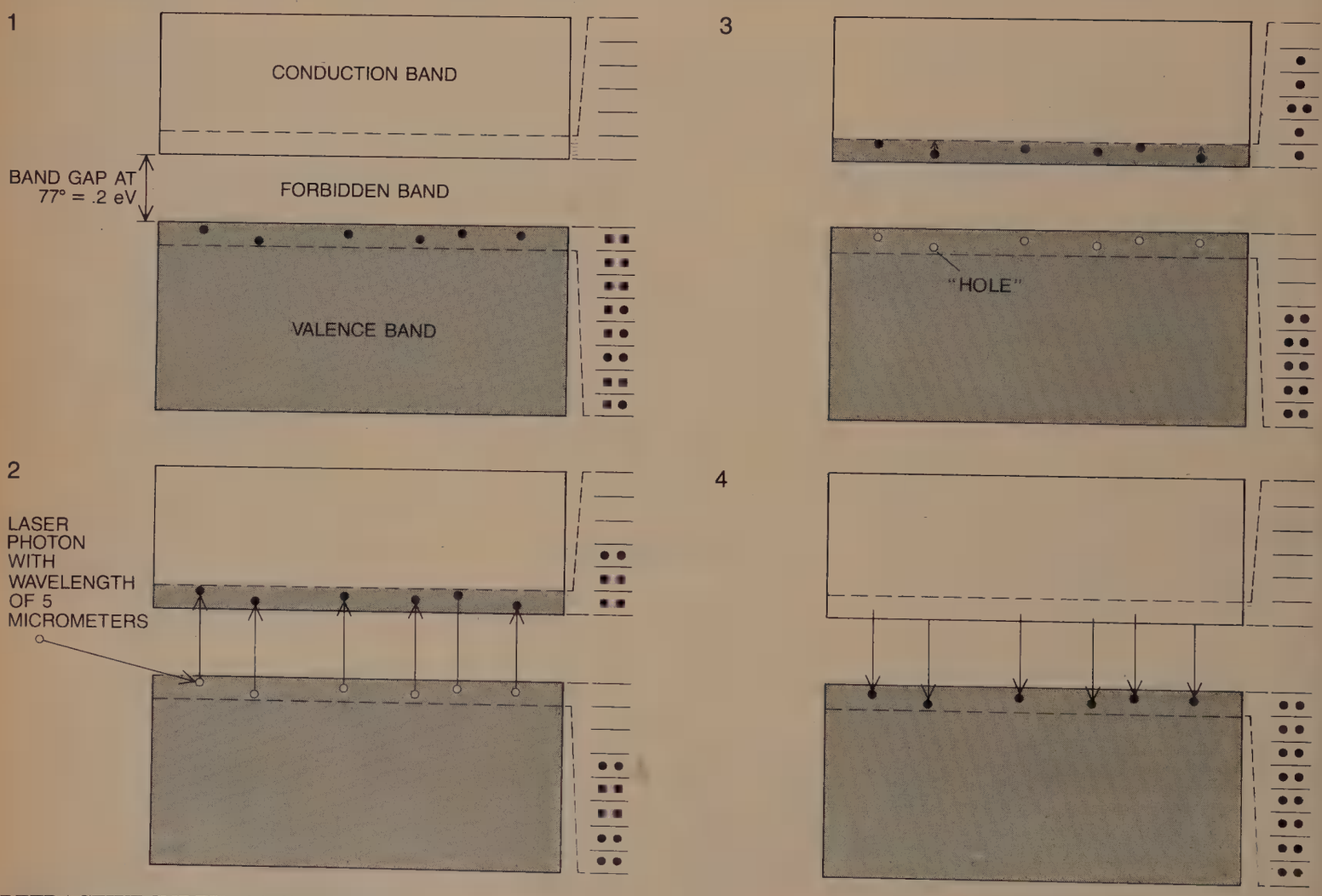
A NOT gate can be constructed by utilizing the reflected beam as the output. Since the reflected beam is the inverse of the transmitted one, increasing the incident intensity will yield low output. Decreasing the incident beam will yield high output.

Toward Photonic Computing

To build an optical computer many optical circuit elements would have to be assembled and interconnected. In the 1970's considerable work was done with thin films through which laser radiation can be made to propagate. The waveguides could in principle form the connecting elements in integrated optical circuits, although the fabrication of such circuits has not yet been attempted.

One application of the new optical technology would be the construction of an optical computer with an architecture, or logical organization, much like that of the standard electronic machine. The chief value of a computer with a conventional architecture but optical components would be its great speed. Optical logic gates could carry out 1,000 billion logical operations per second compared with a maximum of one billion for electronic switching devices.

Optical switches have properties that could also make them suitable for new computer designs. Whereas several currents in a single transistor inevitably become mixed, multiple beams can be passed through an interferometer and still remain separate. It would therefore be possible to utilize a single crystal for several separate switching operations. There could be, say, five laser paths through a crystal, each path forming the site of one operation. With the selection of the appropriate beam intensities each operation could be different. For example, two paths might function as AND



REFRACTIVE INDEX of indium antimonide changes in response to changes in the intensity of radiation passing through it. Indium antimonide is the semiconductor utilized by the authors as the nonlinear material in the optical transistor. In a semiconductor crystal the energy levels available to an electron are arranged in bands; within a band there are many closely spaced levels, each of which can be occupied by two electrons, but energies in the gaps between the bands are forbidden. Here two bands are shown: at low temperature the lower band (called the valence band) is full and the upper band (the conduction band) is empty (*1*). To go from the valence band to the conduction band in the process called promotion an electron must acquire energy equivalent to that represented by the band gap; in indium antimonide the energy is about .2 electron volt. The energy can be acquired by absorption of a photon with a wavelength of five micrometers, which is in the infrared. The authors employed a carbon monoxide laser with an adjustable wavelength to supply such photons (*2*). When an electron absorbs a photon and is promoted to the conduction band, it leaves behind it in the valence band a "hole" that acts like a positively charged particle. Thermal and vibrational energy quickly scatter electrons among the lower levels of the conduction band (*3*). When these levels are filled, no more photons with a wavelength of five micrometers can be absorbed. The blocking of the absorption leads to a concomitant change in the refractive index of the crystal. When the incident beam is switched off, the electrons return to the valence band and recombine with holes, which are thereby eliminated (*4*). The refractive index returns to its original value.

gates, two as OR gates and one path might work in the basic transistor mode.

If the modulated beams for each of the five paths were derived from a single signal that had been divided into five parts, the original signal could be subjected to five logical operations at the same time. The resulting signals could then be processed further, either separately or together. Such a mode of operation would require computer circuits quite different from those of an electronic machine.

Optical switching could lead to other modifications of computer design and logic. One possibility, as we noted above, is the introduction of devices with more than two logical states. Some of the crystals we work with have the capacity to yield successively higher levels of transmitted intensity with successive increases in the intensity of the incident beam. The levels of transmitted intensity might represent the states of a many-valued logic system.

Optical Data Processing

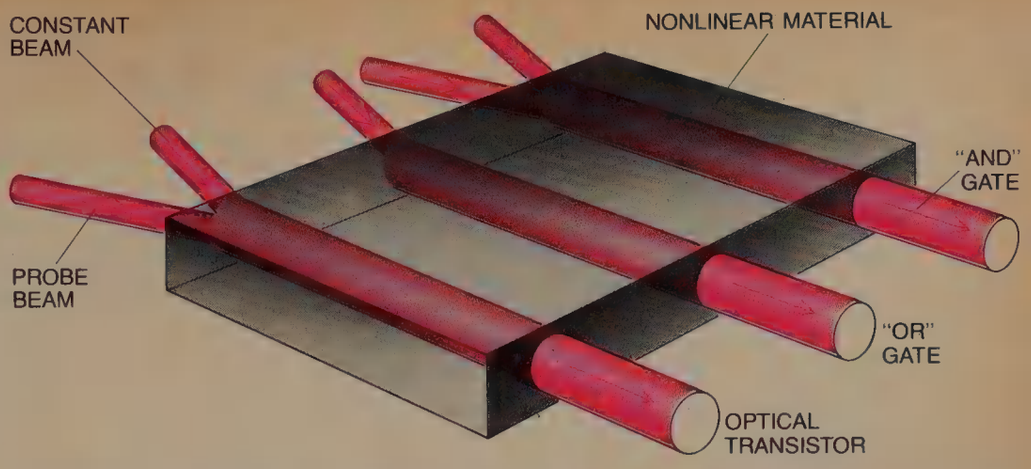
Optical switching systems need not be thought of merely as rapid substitutes for electronic devices. On the contrary, the greatest benefits of optical switches could come from applications that cannot be duplicated by other means. Optical fibers, which have the capacity to carry prodigious amounts of information, are being utilized increasingly for communications, including communications among computers. Optical switching is a natural candidate to mediate between electronic systems and optical ones. On the other hand, if the computation is done optically to begin with, optical fibers could be employed as direct links between computing systems.

As we have seen, optical switches can be made in the form of thin crystals. Current techniques of crystal growing or thin-film fabrication make it possible to fabricate very large thin sheets. If an optical image were projected directly through a large sheet in which each small area served as an optical switch, the transmission from the switches would serve as a digital record of the image. Many kinds of processing and enhancement of the image could be done while the record was being made.

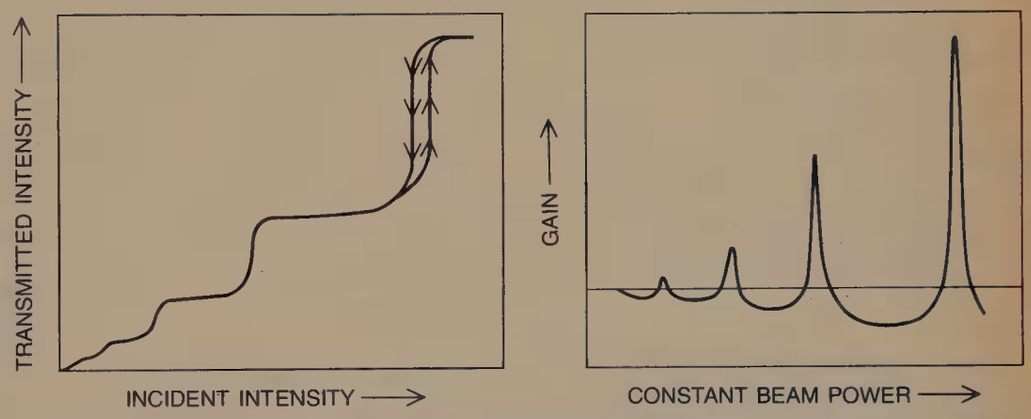
We have expressed considerable enthusiasm about the potential of optical switching devices. It might be well to consider a few of the difficulties that stand between current experimental devices and practical commercial ones. In our own investigations there has been the problem of working at low temperatures. Operation at or near room temperature would obviously make an optical computer more practical. As discussed above, the band-gap energy of indium antimonide is small, and it decreases with increasing temperature. By employing a carbon monoxide laser with a wavelength of 10 micrometers, we have recently been able to achieve optical switching at room temperature.

New lasers and new nonlinear materials will undoubtedly make it possible to improve the speed and efficiency of optical switches. The question of appropriate materials is at the center of the practical problems of building an optical computing machine. In general it is desirable that a switch consume as little power as possible. The switching power is the product of the light intensity incident on the crystal and the area of the crystal face. Reducing the size of the device can therefore reduce the switching power. The device cannot be made too small, however, because there is a limit to how finely a laser beam can be focused. In addition, as the size of the device decreases overheating becomes a problem.

Balanced against the general desirability of achieving a low switching power is the fact that within certain limits the switching time can be reduced by applying more power. The size of the device, the switching speed and the switching power must all be considered in relation to one other, to the optical and thermal properties of the nonlinear medium and to the function of the switch. The selection of the right medium or the discovery of new nonlinear materials is probably the most fundamental task in the development of optical switching. In spite of such formidable difficulties we think the optical computer is a realistic and exciting prospect.



PARALLEL PROCESSING is a capacity of optical switches that could lead to new designs and new capabilities in the computer. Multiple laser beams can be focused so that they remain separate as they pass through a crystal of nonlinear material. Each radiation path could serve as the site of a separate operation, and the operations carried out on the various beams could be different. For example, if there were three paths, one could be an AND gate, one an OR gate and one a transistor. If three incident beams came from one beam that had been split, three operations at once could be done on the original signal. Such a capacity would require a new form of information processing; in electronic computers one operation at a time is done on a signal.



MULTIPLE BISTABILITY is a property of optical switches that could lead to the development of a new system of computer logic. For a given interferometer there are many optical path lengths equal to an integral multiple of the half wavelength of the radiation. Each such optical length corresponds to a peak in the Airy function and so yields maximum transmission. Increasing the intensity of the incident beam causes the optical length of the nonlinear crystal in the cavity to reach successive peaks. The graph of transmitted intensity plotted against incident intensity thus shows a series of nearly vertical steps, each step corresponding to a different peak in the Airy curve (*left*). If the intensity of the constant beam in the optical transistor is increased, successive operations of the smaller probe beam take the transmitted power up successive vertical steps in the curve of transmitted intensity. Since the vertical steps in the curve at the left are progressively greater, the gain, or amplification, of the output by the operation of the probe beam increases with each switching operation (*right*). The different transmission levels could be made to stand for the values of a logic system that takes the place of binary logic.

Hidden Visual Processes

Vision is usually regarded as being a single sense. Experiments show, however, that the visual system includes subsystems whose operation is normally hidden from the awareness of the perceiver

by Jeremy M. Wolfe

From the standpoint of a person contemplating his abilities to perceive the world, vision seems to be a single sense. The images impinging on the two retinas give rise to a single awareness of the objects in the world: their sizes, shapes, colors, textures and positions. This view, however, is mistaken. The visual system (the brain mechanism that processes data from the eyes) is actually a set of specialized subsystems each of which acts more or less independently on some subset of visual data. Furthermore, some of the visual subsystems have an output that cannot be seen. They contribute to brain function and even to our awareness of the world, but no amount of introspection can make us aware of the subsystems themselves. They perform hidden visual processes.

How is a hidden process revealed? One way is to examine the abilities of people who have suffered brain injury. Consider the pupil of the eye, which constricts in response to an increase in the intensity of light falling on the retina. If someone suffers an injury that destroys the visual cortex (the part of the cerebral cortex first in line to get data from the eyes), he is rendered perceptually blind. That is, his awareness of a loss of vision is the same as that of someone who has lost the use of the eyes themselves. Nevertheless, the pupils continue to constrict in response to light. Even more strikingly, Ernest C. Poppel, Richard Held and Douglas Frost of the Massachusetts Institute of Technology have found that when people who are perceptually blind because of an injury to the cerebral cortex are asked to direct their eyes toward a spot of light, they do surprisingly well. The subjects report that they cannot see the spot, and so they think they are guessing, but they look in roughly the right direction more often than chance would allow.

Brain-injured people thus show evidence of multiple visual processes: some that are damaged, and hidden ones that remain functional. What such studies cannot confirm is that similar hidden processes operate in people whose visual system is intact. For that, special experimental strategies are required. Here I shall describe three sets of experiments each of which reveals a visual process that in normal people is hidden from introspection. In this regard all of us are like the brain-injured patient who can look at a spot of light although he cannot perceive it. We too are unaware of the abilities of certain parts of the visual system even though our behavior is often based on their output.

One of the functions of the visual system is to control the muscles that focus the eye on objects at various distances by changing the shape of the lens. The closer the object is, the more nearly spherical the lens must be. The process is called visual accommodation. One's impression is that one can accommodate for anything one can see. It is natural to assume, therefore, that accommodation and visual perception have access to the same set of stimuli.

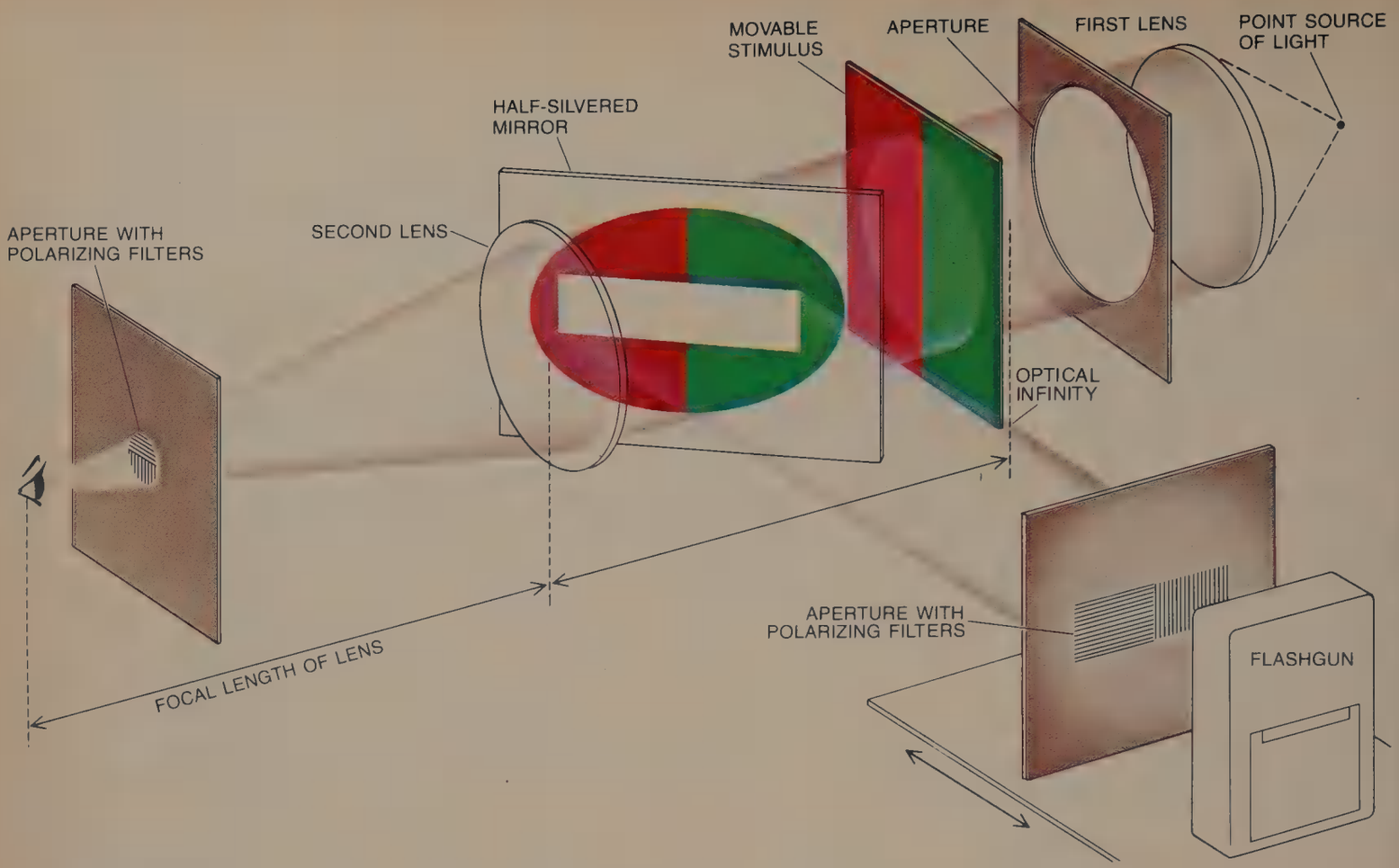
Do they? By means of a number of experimental methods the accommodative status of the eye can be measured while the subject looks at a stimulus placed a certain distance from him. In one such method the subject views the stimulus through polarizing filters, and a flash of light from behind a slit briefly superposes on the stimulus a bright horizontal bar. The filters have no effect on the stimulus, but they ensure that light from the left half of the bar will enter only the top of the lens of the subject's eye and light from the right half will enter only the bottom.

Suppose a subject who is looking at the stimulus is accommodating for a distance greater or less than the distance to the slit. If the bar is flashed at that instant, the two halves of it will be misaligned on the retina and the subject will see them as being offset. Hence the experimenter need only have the subject look at the stimulus while the bar repeatedly appears and ask the subject whether or not its halves are aligned. The slit is moved to various distances until the subject reports that they are aligned. The distance from the slit to the eye is then the distance for which the subject is accommodating.

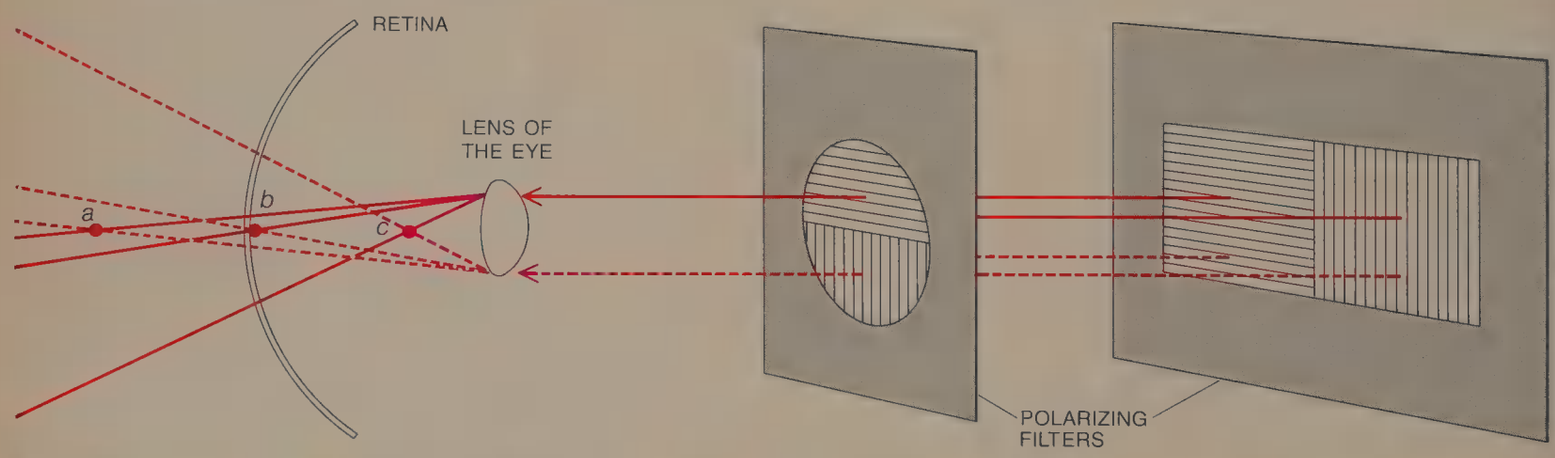
The lens of the eye hardens with age, and so it is best to study accommodation in subjects no older than their 30's. In the laboratory such subjects can be shown stimuli that change only in distance, not in brightness and size. This eliminates all perceptual clues to the distance of the stimulus except the fact that a certain accommodative state brings the stimulus to a focus. Under these circumstances the typical subject's accommodation is about 90 percent of perfect. (If it were perfect, a stimulus at a distance of 25 centimeters would cause the eye to accommodate for a distance of 25 centimeters and so on.)

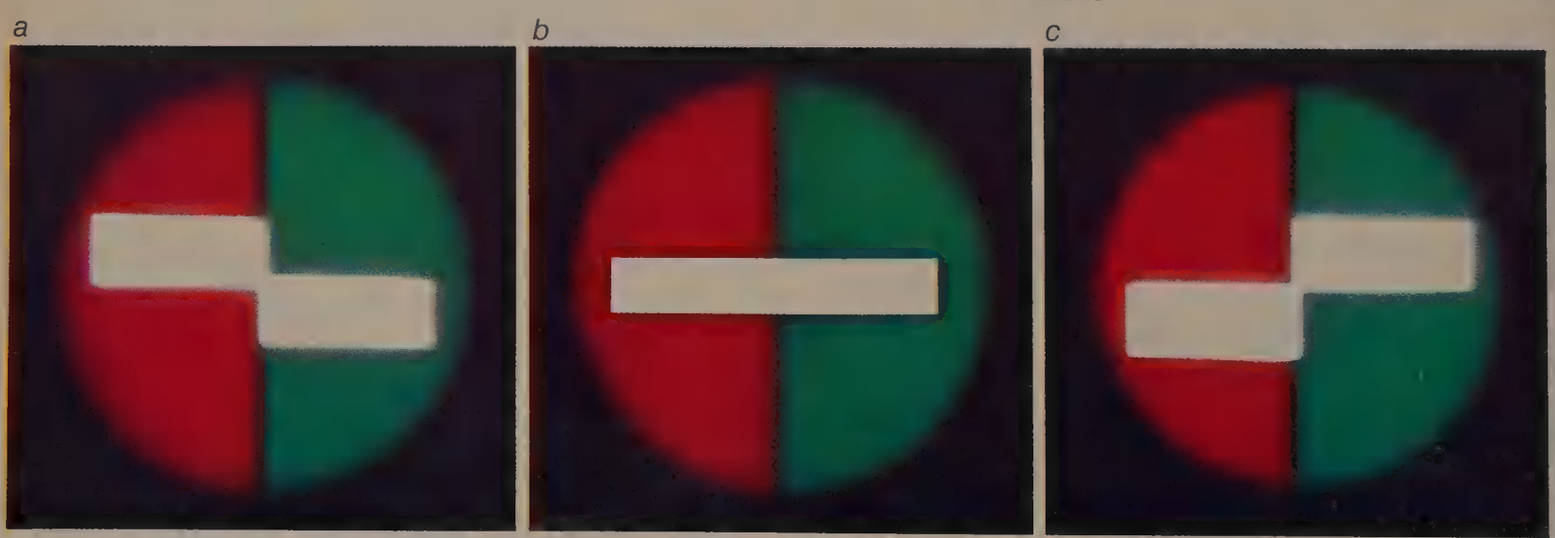
How does the eye respond when the stimulus has no features for the eye to focus on? One such stimulus would be a blank screen surrounding the subject; the experience is like being inside a giant ping-pong ball. Another stimulus would be a smaller blank screen viewed through a lens that makes the screen appear too close (in optical, or apparent, distance) for the eye to focus

ISOLUMINANT STIMULUS is an image whose edges are defined only by a change in color, not by a change in brightness. The stimulus here is imperfect: the blue parts and the green parts of the image are only as nearly equal in brightness as they can be on the printed page. Moreover, the change in brightness beyond the edge of the page is apparent, and so is the fact that the reader is holding the magazine at reading distance. When such cues are removed under laboratory conditions, subjects faced with an isoluminant stimulus prove unable to bring its edges into focus. This deficiency contributes to making a familiar face hard to recognize. The experiment indicates that the brain process underlying visual accommodation (the focusing of the eyes) cannot "see" color; it is a hidden process distinct from the processes that lead to perception. The image shows Groucho Marx as he appeared in the motion picture *Horse Feathers*.



EXPERIMENTAL APPARATUS tests the ability of the eye to accommodate for an edge in an isoluminant stimulus. The lens toward the upper right in the illustration collimates a beam of light. The beam passes through an aperture, then through the stimulus, which has a vertical edge defined only by the colors red and green. A second lens directs light into the subject's eye. A further arrangement consisting of a flashgun, a movable slit and a half-silvered mirror briefly superposes the isoluminant stimulus a horizontal bar of light.

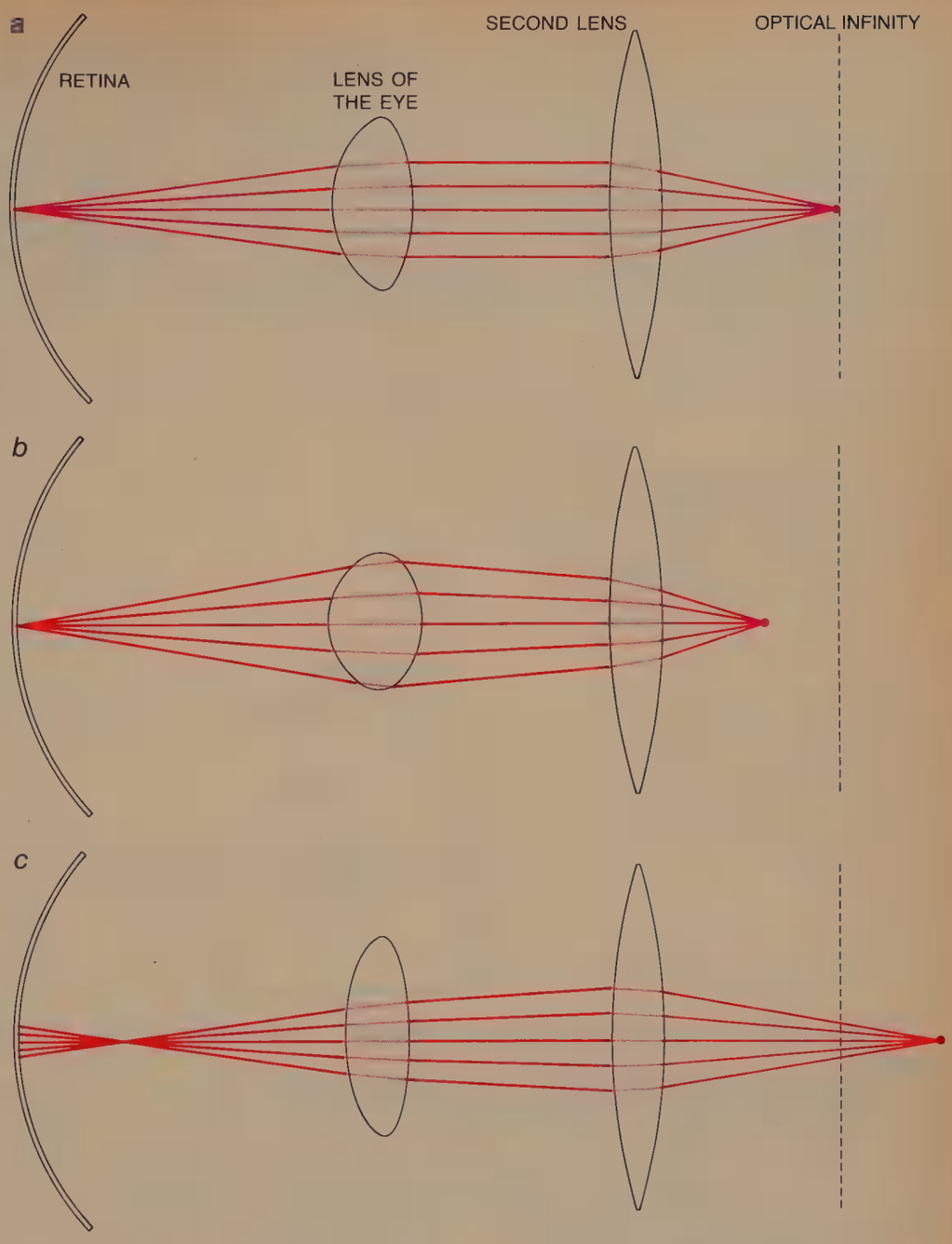




ACCOMMODATION IS MEASURED by having people describe the flashing horizontal bar. A series of polarizing filters ensures that light from the left side of the bar enters only the top of the lens of the subject's eye; light from the right side enters only the bottom. The two sides of the bar will line up on the retina (*b*) only if the subject happens to be accommodating for the distance to the movable slit.

on its edges. Another would be a featureless sky; still another would be complete darkness. Herschel W. Leibowitz and D. Alfred Owens of Pennsylvania State University have found that in any of these circumstances the lens of the eye assumes a rather stable resting curvature for a particular focal distance called the dark focus. Each individual has a characteristic dark focus that is usually about one meter, or about an arm's length from the eye.

The same thing happens when the stimulus is a grating of black and white lines so fine that they give the impression of a gray field. Here again the experimenter's lens can place the stimulus at a variety of optical distances. When the stimulus is at an optical distance of 25 centimeters, the lens of the subject's eye assumes its resting state. When the stimulus is moved to an optical distance of one meter, the subject's accommodative state shows no change. The conclusion is that accommodation can shape the lens of the eye to keep an object in focus only if the system responsible for accommodation is presented with something it can "see."



OPTICAL DISTANCE of a stimulus is the apparent distance given the stimulus by the second lens in the experimental apparatus. If the stimulus is positioned at the focal length of the lens (*a*), it will be at optical infinity, that is, the light rays from each part of the stimulus will approach the eye in parallel, as if they came from a source infinitely far away. The lens of the eye will assume a curvature that brings the rays to a point on the retina. If the stimulus is closer than optical infinity (*b*), the rays will diverge as they approach the eye. The lens of the eye will then assume a more nearly spherical shape to bring the rays to a focus. If the stimulus is farther away than optical infinity (*c*), the rays will converge as they approach the eye. The lens of the eye will be unable to reduce its curvature to the required extent; thus the rays will come to a point in front of the retina. A stimulus beyond optical infinity cannot be brought into focus.

Armed with the knowledge that the accommodation system is blind to certain aspects of the visual world, Owens and I asked whether the system can "see" color. We measured accommodation while our subjects looked at a simple stimulus: a circular field divided vertically in half so that it included a single vertical edge. An optical system known as Maxwellian-view optics ensured that the perimeter of the circle was beyond optical infinity. That is, the rays of light from each point on the perimeter of the circle were made to enter the eye in such a way that they could never be brought into focus whatever shape the lens of the eye assumed. As a result the vertical edge between the two halves of the field was really the only edge in the stimulus for which the eye could accommodate. The optical system was also designed to ensure that the image of the stimulus would keep the same size on the retina when we changed its optical distance.

The edge itself could be created by a difference in either color or brightness between the two half fields. In the real world most contours arise from a difference in both. We chose to make one half field red, the other green. Then we varied the brightness of one half field or the other. In this way we created stimuli ranging from an edge between red and black to an edge between black and green. At each extreme the contrast in brightness across the edge between the color and the black was 100 percent. In the precisely intermediate case, however, the red and the green were equally bright. They formed an isoluminant stimulus whose single edge was defined only by color. The contrast in brightness was zero.

When we tested people's ability to accommodate for each of these stimuli, we found the ability declined as the contrast in brightness declined. Thus our subjects could readily tell that the isoluminant stimulus was an edge between red and green, but it was impossible for them to bring the edge into focus. In one experiment we had subjects look at a black *E* on a white background. The subjects accommodated quite well for optical distances from infinity down to 22 centimeters. Then we had them look at edges between red and black or between black and green. Their accommodation was about 80 percent as good as the best they had done for the *E*. Finally we showed them the red-green edge that has no contrast in brightness. Their accommodation was only 19 percent as good. When we repeated the experiment with pairs of colors such as red and orange or blue and green, the subjects' performance was equally poor or worse.

Like any other optical system, the eye has a chromatic aberration, which brings different colors to a focus at slightly different focal lengths. The aberration therefore tends to shift the retinal image of one color half field with respect to the image of the other. This can give rise to a bright or dark contour

Rover's family tree begins 200,000,000 years ago.

The award-winning creator of *Archosauria* and *Synapsida* now explores the natural history, variety, and adaptability of the dog, from the prehistoric ancestors he shared with man to present-day wild canids and *Canis familiaris*. "Readers are sure to emerge with a deeper understanding of just how and why this adaptive creature became man's best friend.... McLaughlin's most broadly appealing book yet."—*Kirkus Reviews*

87 illustrations

THE CANINE CLAN

A New Look at Man's Best Friend

JOHN C. McLOUGHLIN

THE VIKING PRESS Dept. ATE-SA
40 West 23rd St., New York 10010
Send ___copies of THE CANINE CLAN @ $15.75. Also send ___copies of *Archosauria* and ___copies of *Synapsida* @ $14.95 each. I enclose $_____ total. (Add sales tax where applicable plus $1.75 per copy for shipping.)

Name_____

Address_____

City_____ State_____ Zip_____

Garden Camera CALCULATORS

Hewlett Packard

HP 67	$288.50
HP 97	558.50
HP 32E	48.50
HP 37E	57.95
HP 33C	63.50
HP 34C	74.50
HP 38C	74.50
HP 41C	148.50
HP 41CV	207.50
Accessories	Call
Printer	Call
Card Reader	Call
HP 11C	74.50
HP 12C	111.50
HP 15C	99.50
HP 16C	114.50
HP 10C	59.50

Texas Instruments

TI 58C	78.95
TI 59	168.50
PC 100C	149.50
TI MBA	49.50

Criterion Telescope 499.00

master charge VISA

Call for low prices on Nikon, Minolta, Olympus and all Major Brand Cameras

Prices subject to change without notice

Call 1 (800) 223-0595

Or Send postage and handling to

GARDEN CAMERA

135 West 29 Street, N.Y., N.Y. 10001
Or Visit Our Store: 345 Seventh Avenue
New York, Alaska & Hawaii Call:
Tel: (212) 868-1420—Open Weekdays 8:30–6:00
OPEN SUNDAYS 10–4 p.m. Closed Saturdays

Speak German like a Diplomat!

What sort of people need to learn a foreign language as quickly and effectively as possible? *Foreign service personnel,* that's who.

Now you can learn to speak German with the same materials used by the U.S. State Department — the Foreign Service Institute's *Programmed Introduction to German.*

The FSI spent thousands of dollars and many years developing these materials for use by members of the United States diplomatic corps. Today people in all walks of life who need to learn to speak a foreign language are turning to this outstanding audio program.

The FSI's German Course is by far the most effective way to learn German at your own convenience and pace. It consists of a series of cassettes, recorded by native German speakers, and accompanying textbook. You simply follow the spoken and written instructions, listening and learning. By the end of the course you'll find yourself learning and speaking entirely in German!

This course turns your cassette player into a "teaching machine." With its unique "programmed" learning method, you set your own pace—testing yourself, correcting errors, reinforcing accurate responses.

This Programmed Course comes in two volumes, each shipped in a handsome library binder. You may order one or both:

☐ **Volume I. Programmed Introduction to German,** 9 cassettes (12 hr.), and 647-p. text, $115.

☐ **Volume II. Basic Course Continued,** 8 cassettes (8½ hr.), and 333-p. text, $98.

(Conn. and N.Y. residents add sales tax.)

TO ORDER BY PHONE, PLEASE CALL TOLL-FREE NUMBER: **1-800-243-1234.**

To order by mail, clip this ad and send with your name and address, and a check or money order — or charge to your credit card (AmEx, VISA, MasterCard, Diners) by enclosing card number, expiration date, and your signature.

The Foreign Service Institute's German course is unconditionally guaranteed. Try it for three weeks. If you're not convinced it's the fastest, easiest, most painless way to learn German, return it and we'll refund every penny you paid. Order Today!

81 courses in 26 other languages also available. Write us for free catalog. Our 10th year.

Audio-Forum
Suite 23-D
On-the-Green,
Guilford, CT. 06437
(203) 453-9794

AUDIO-FORUM®

Or visit our New York sales office: 145 E. 49th St., New York, N.Y. 10017 (212) 753-1783

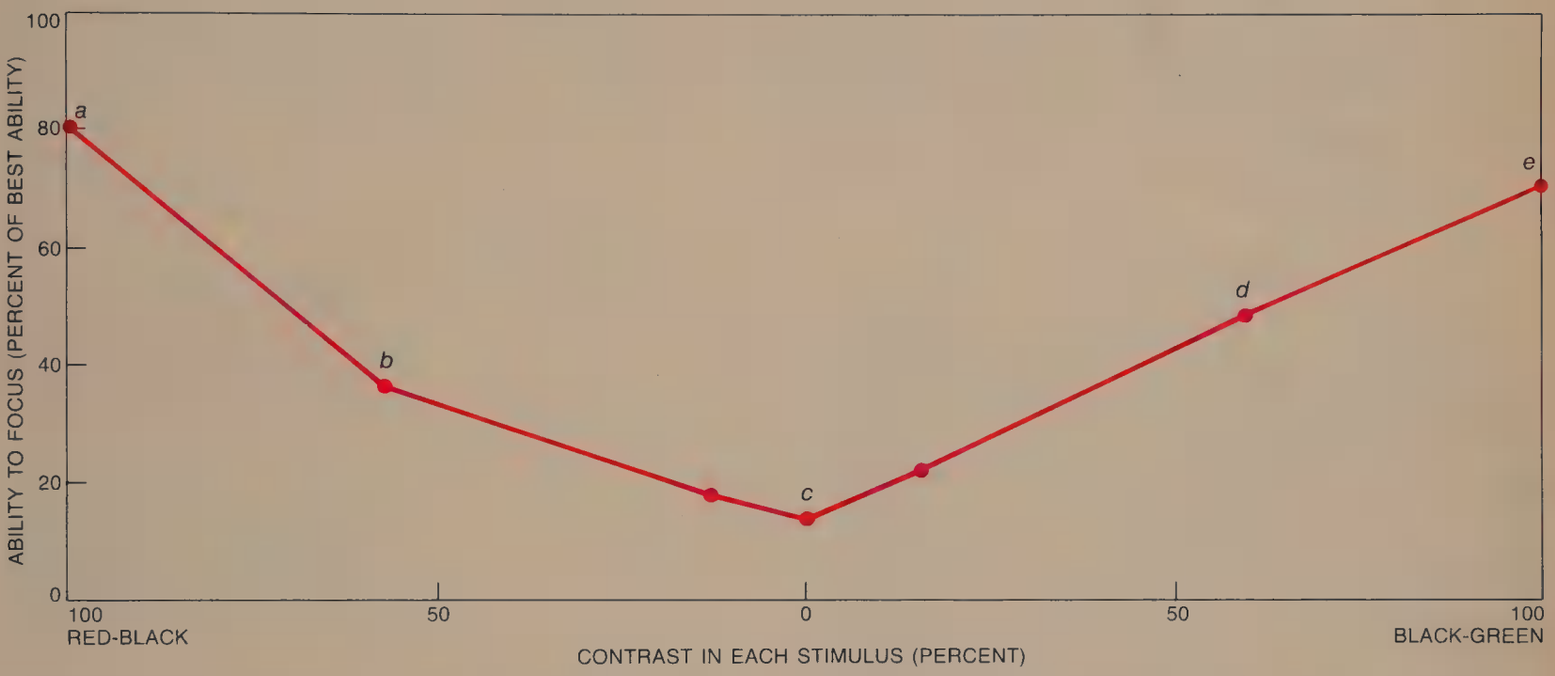
depending on whether the images separate or overlap. As a result it is surprisingly difficult to create a stimulus that has absolutely no clues to the presence of an edge except the difference in color itself. We suspect our subjects' performance would have been even worse than it was if chromatic aberration could have been eliminated. Still, the conclusion seems clear: accommodation is colorblind, or at least it "sees" color very badly. It is a visual subsystem remarkably independent of visual perception. Here, then, is a finding that no amount of introspection could have suggested about vision.

A second strategy for exploring the division of the visual system into subsystems is to consider the ways a particular visual function is performed by different parts of the system. One such function is binocular vision. A normal person has two eyes, and the brain makes every effort to integrate the data from them. It is becoming clear, however, that the brain does more than simply combine the inputs from the eyes in one grand binocular process that leads to visual perception. Instead several special-purpose mechanisms combine the inputs in their own way to meet their own particular needs. Just as no amount of introspection will reveal that accommodation is colorblind, so no amount of introspection will reveal these multiple binocular processes. Nevertheless, the processes do exist and can be revealed by experimentation.

Consider a visual illusion that everyone has experienced in one form or another. You are sitting in a train, waiting for it to pull out of the station. Looking out the window, you see another train motionless on the adjacent track. The other train's image starts to slide backward, and you distinctly feel that your train is moving forward. Then you see it is the other train that has moved; your train is still in the station. The motion of the other train somehow deluded you into believing you were moving. You have experienced an illusion of motion created solely by visual stimulation. The illusion is known as vection.

Several investigators have sought to examine the brain mechanism responsible for vection. Held and I became interested in a rather different question. We wanted to know what the vection process can "see." In particular we wanted to know if the process can make use of people's ability to look at the world through two eyes. One might think the answer could be found quite easily by comparing the vection experienced when both eyes are open with the vection experienced when only one eye is open. Normally, however, the sensation of vection is already at its maximum when only one eye is open; the illusion of self-motion cannot be more pronounced. One

STIMULI FOR THE EXPERIMENT vary the ability to accommodate vary from an edge between red and black (*a*) to an edge between black and green (*e*). The precisely intermediate stimulus (*c*) is isoluminant. The circular perimeter of each stimulus is defined by an aperture placed well beyond optical infinity in the experimental apparatus; hence the subject's eye cannot bring the perimeter into focus.



ABILITY TO FOCUS on the stimuli turns out to depend on their contrast in brightness. Here the average performance of four subjects is charted as a percentage of the ability they had when they looked at black letters on an eye-examination chart. The subjects' performance grows poorer with decreasing contrast. It is worst (19 percent of the best performance) when the stimulus is isoluminant.

needs, therefore, to create a stimulus whose binocular effect could conceivably be greater than its monocular effect. Such a stimulus must have a purely binocular component: there must be something about the stimulus that is invisible to each eye by itself.

A component that meets this requirement is called a cyclopean stimulus, after the Cyclops, the Homeric one-eyed creature encountered by Odysseus. The name is apt because a cyclopean stimulus is evident only if the brain combines the input from the eyes. In effect the brain must act as a single, cyclopean eye. One example of a cyclopean stimulus is the small difference in position between the image of an object in the world on each of the retinas. The difference can easily be seen. Stretch out one of your arms and look at the tip of a finger first with one eye closed and then with the other eye closed. You will notice that the two views are slightly different. The brain exploits such differences in the processes that lead to the perception of three-dimensional depth. Clearly the brain must draw on data from both eyes for the stimulus to exist.

Our cyclopean stimulus for vection capitalized on the well-known fact that objects seen in a motion picture seem to move smoothly even though they are presented in a succession of still photographs. Our subjects sat inside a cylinder three feet in diameter and five feet high. The inside surface of the cylinder was white and was covered with a random pattern of black dots an inch in diameter. The cylinder rotated about the subject at the rate of 30 degrees of angle per second. When the inner surface of the cylinder was illuminated by ordinary lighting, the subject reported a sensation of rotation in the opposite direction. For our experiments we illuminated the surface with the periodic flashes of a stroboscopic lamp. Each flash produced the equivalent of a frame from a movie of black dots. The subject reported the same sensation of vection.

To create a cyclopean stimulus two strobe lamps were needed. One was covered with a red filter, the other with a green filter. The subjects wore goggles that placed a red filter in front of one eye and a green filter in front of the other. No red light could pass through the green filter; no green light could pass through the red. Hence the light from one strobe lamp was seen by one eye and the light from the other strobe lamp was seen by the other eye.

Suppose each strobe lamp is flashing at 10 hertz, or 10 times per second. If the two lamps are flashing in phase (that is, in synchrony), the subject should notice no important difference in looking with two eyes rather than one. In either case he will see in effect a 10-frame-per-second movie of moving dots. Suppose, however, the strobe lamps are exactly out of phase, so that they flash in alternation. Now there is a difference. A subject who has only one eye open will see again a 10-frame-per-second movie. With both eyes open he will see an additional movie. It will appear at a rate of 20 frames per second, and it will consist of frames presented in alternation first to one eye, then to the other, then to the first eye again.

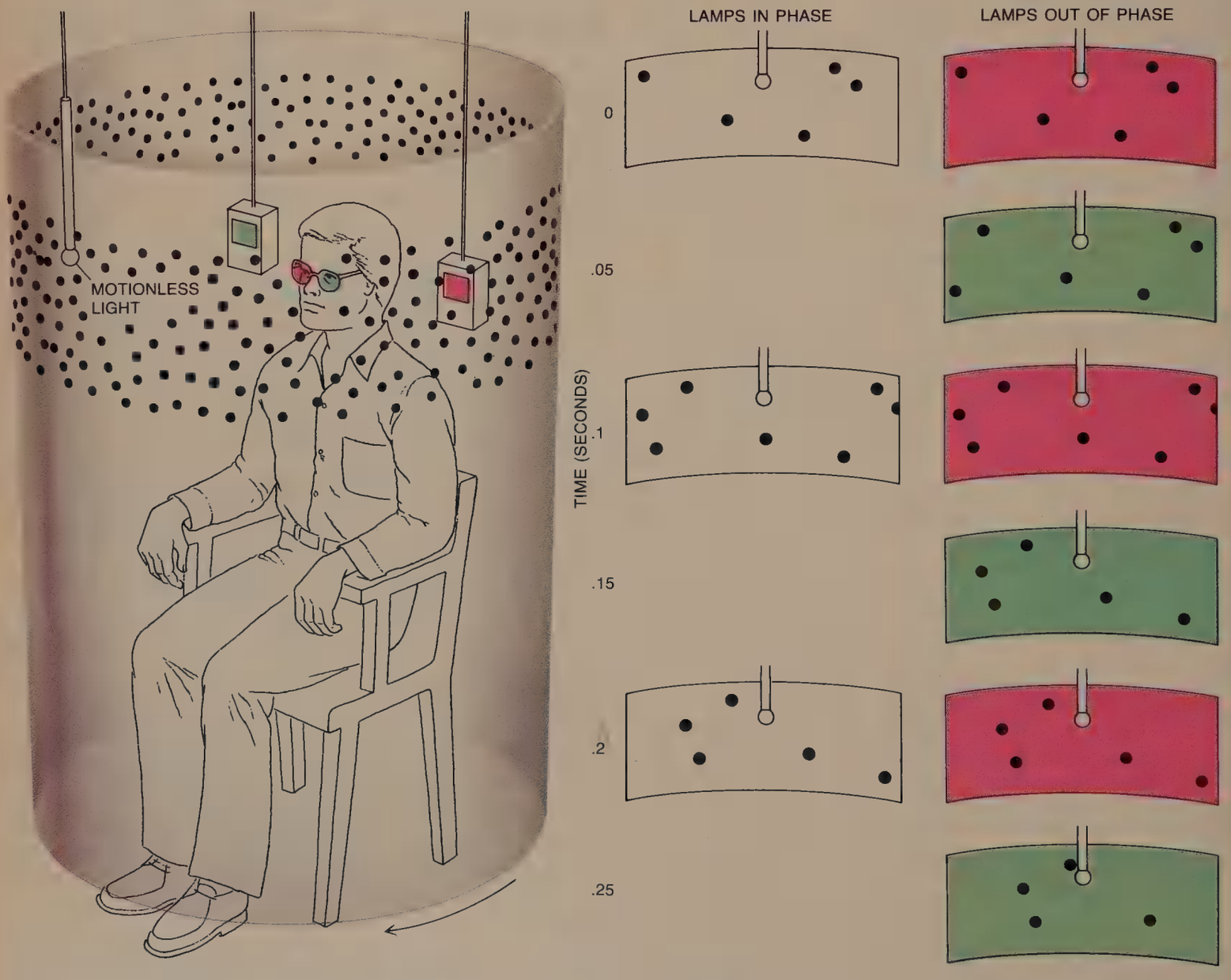
The 20-hertz interocular movie is thus

a cyclopean stimulus: it cannot be seen by either eye alone. It does give rise to the appearance of motion. The question is whether the vection system can "see" it. If it can, the experiment where the strobe lamps are out of phase could produce a greater sensation of vection than the experiment where the lamps are in phase. If it cannot, the two experiments should have the same result.

With our apparatus in place we prepared subjects for the experiments. In particular we taught them a method known in experimental psychology as magnitude estimation. We asked them to give a rating of 10 to a compelling sensation of self-motion and one of zero to no sensation of motion. Ratings between 0 and 10 were to be given to sensations between these two extremes. Unlikely as it may seem, some 25 years of research in experimental psychology, notably the work of S. Smith Stevens of Harvard University, has shown that people are quite good at assigning numbers to their sensations. They can do it repeatedly, systematically and reliably.

One thing did concern us: there could be no difference between the results of the two experiments if the sensation of vection yielded the maximum rating of 10 when the lamps were in phase. If the frames of a movie are shown at a progressively lower rate, the illusion of motion becomes progressively less compelling. In the same way the flicker rate turns out to be important to the sensation of self-motion. We found that any rate higher than about 2.5 hertz produced some degree of vection, but only rates higher than about 15 hertz consistently earned ratings of 10 from our subjects. Therefore flicker rates between 2.5 hertz and 15 hertz would suit our purpose. The cyclopean stimulus created (at twice the flicker rate) by flashing the lamps out of phase would be capable in principle of producing a more compelling sensation of vection than the stimulus produced by leaving the lamps in phase.

Our findings were straightforward. For any flicker rate between 2.5 hertz and 15 hertz the two strobes out of phase always elicited the greater magnitude estimates. Evidently the vection system can "see" the purely binocular stimulus. Does this result establish that the brain has multiple binocular processes? Not in itself. As I have noted, the slight difference in position between the images of an object on each retina is



EXPERIMENT WITH VECTION (the sensation of bodily motion arising from the motion of the visual world) requires that subjects be seated inside a rotating cylinder whose inner surface is white and is covered with a random pattern of black dots. Two stroboscopic lamps illuminate the surface in flashes of red and green light; a pair of goggles allows the light of each color to enter one eye only. If the lamps flash in phase (that is, synchronously) at a rate of 10 flashes per second, a subject with one eye open sees what a subject with both eyes open sees: in effect a 10-frame-per-second motion picture of moving dots. If the lamps flash out of phase, a subject with one eye open sees the identical movie, but a subject with both eyes open sees an additional movie. It appears at a rate of 20 frames per second, and it is interocular: its successive frames are seen by the two eyes in alternation. A small stationary light serves as a reference point.

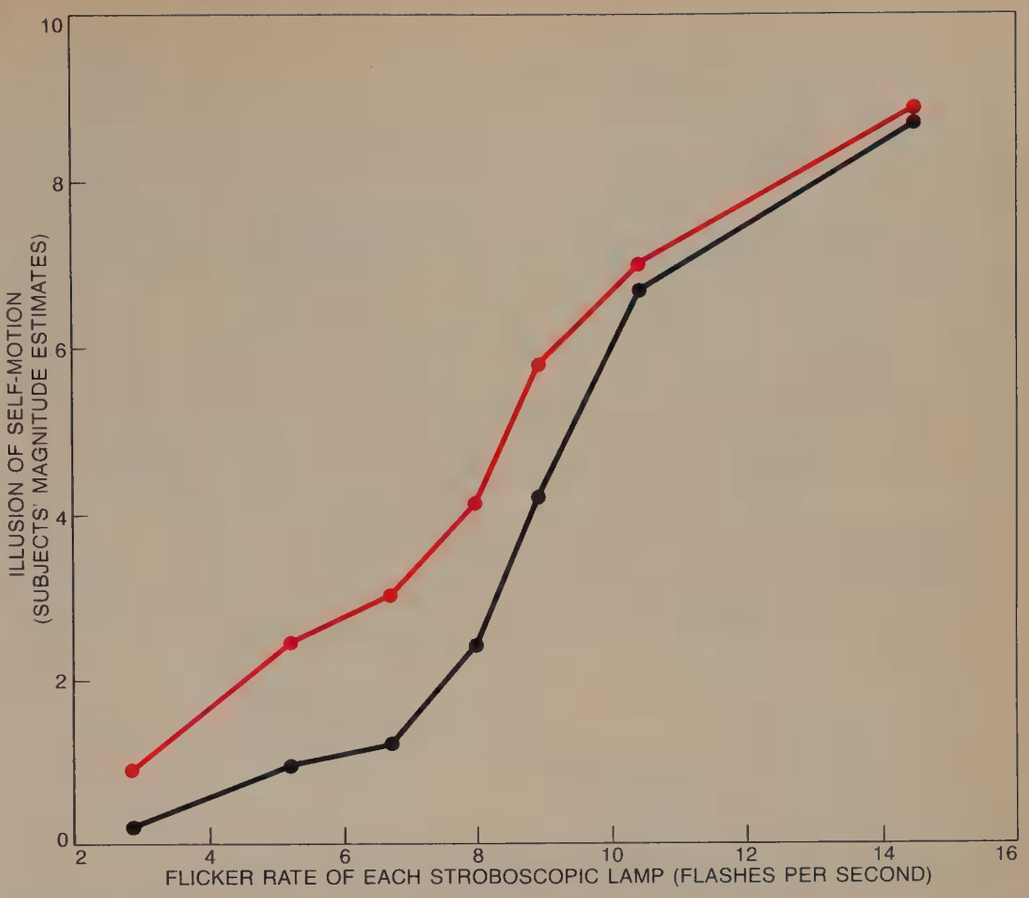
a cyclopean clue to three-dimensional depth. It is conceivable that the binocular mechanism serving this aspect of visual perception also serves the binocular contribution to vection.

Some people, however, cannot perceive depth on the basis of cyclopean stimuli. For example, they cannot get the illusion of depth when they look at an image through a stereoscope or when they go to a "3-D" movie. They are stereoblind. Some of them are born that way, just as some people are born colorblind. Others lose the ability because of defects that develop in infancy in the ability to align the eyes. Stereoblindness is somewhat rarer than colorblindness; it seems to affect only a few percent of the U.S. population. A person who is stereoblind can still perceive depth on the basis of visual cues such as the apparent size of familiar objects or the fact that some objects in the visual field block others that are farther away. These monocular cues to three-dimensional depth are quite good; many people who are stereoblind do not know they are until a test reveals it.

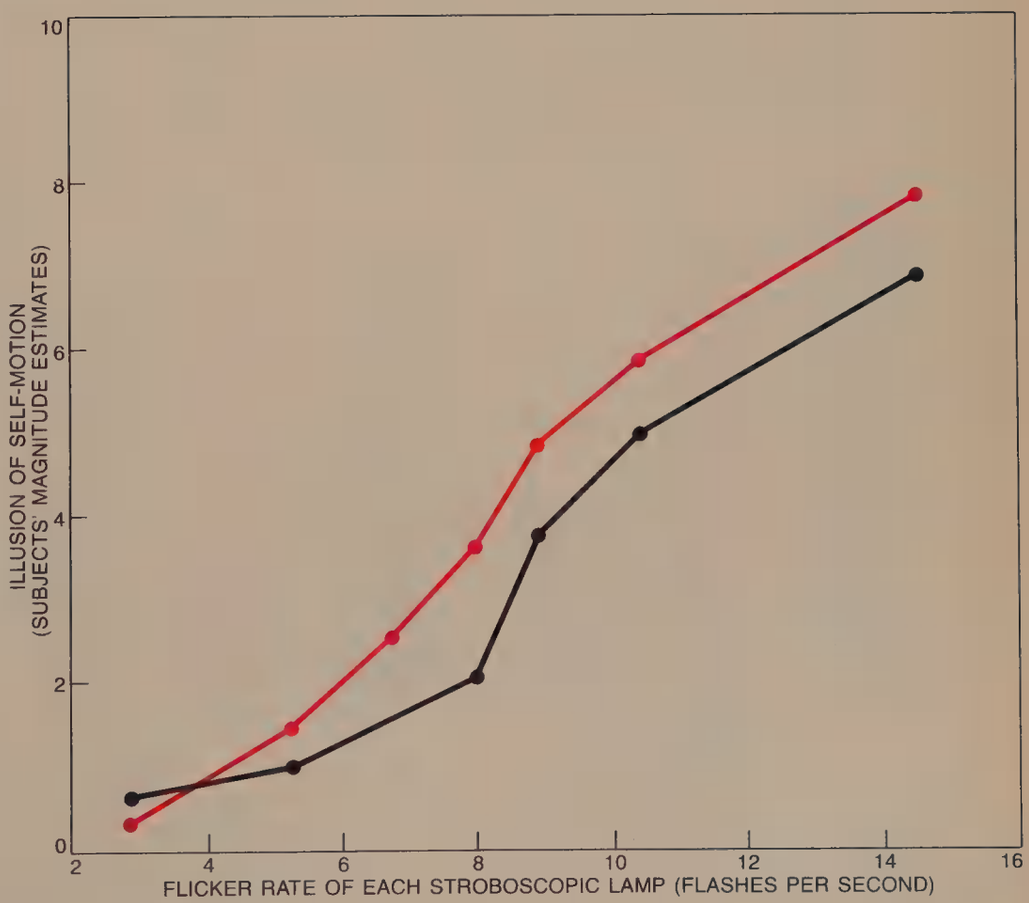
Held and I asked four people who had been established as stereoblind to be subjects in our vection experiments. Their magnitude estimations were much like those of our normal subjects. They too got an increased sensation of self-motion from the purely binocular stimulus. Here, then, are people with a defective binocular mechanism who nonetheless prove to be perfectly normal in experiments requiring that their brain employ a binocular mechanism. Thus the defective mechanism cannot be the only binocular mechanism in the brain. There must be at least two binocular mechanisms, one mediating stereoscopic depth perception, the other involved in the production of vection. The experiments again show that visual data feed more than one processing system.

The two hidden visual processes I have now described (the colorblind visual process that controls accommodation and the binocular visual process that contributes to vection) appear to play no direct role in visual perception. Other hidden visual processes do take part in perception.

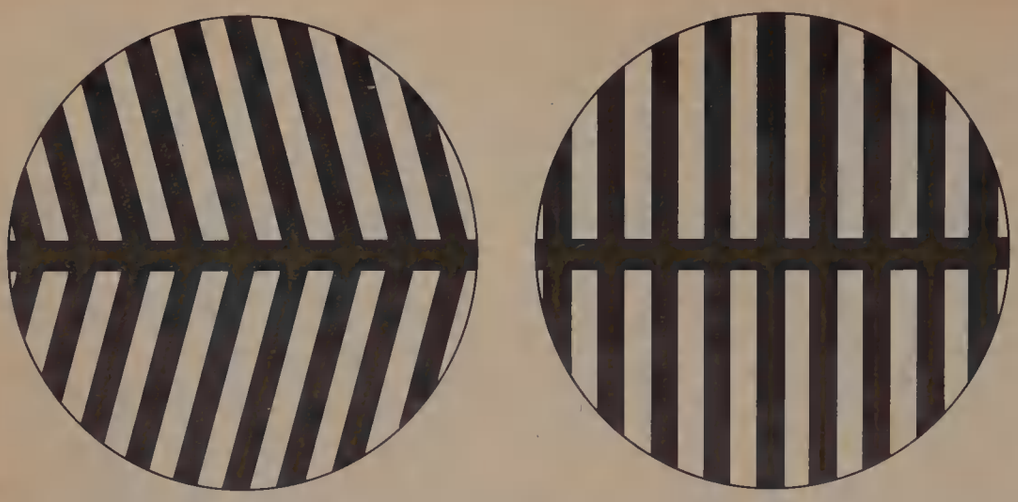
For an example I shall return to the binocular visual processes. Our experiments with vection revealed two binocular processes. The experiments I shall now describe show more. Indeed, there emerges a remarkable assortment of processes. Between our two eyes and our single perception of the visual world lie processes that can "see" out of one eye, processes that can "see" out of either eye and, most surprising, a purely binocular process: a process that can "see" only out of both eyes. In fact, it can "see" only when both eyes are looking at the same stimulus. If your visual



SENSATION OF VECTION is greater if the lamps are out of phase (*color*) than it is if the lamps are in phase (*black*), showing that the brain process responsible for vection "see" a movie created by stimulating the eyes in alternation. It is thus a binocular process. The data in the chart were collected by asking six subjects to assign ratings in which 10 signified the most compelling illusion of bodily motion whereas zero signified the absence of such an illusion.



RATINGS BY STEREOBLIND SUBJECTS show that they too have a greater sensation of vection if the lamps are out of phase. People who are stereoblind lack the binocular visual process that compares the images from the eyes to aid in the perception of three-dimensional depth. Hence the process responsible for vection must be a different binocular process. The data in the chart were collected from four stereoblind subjects, who rated their sensations.



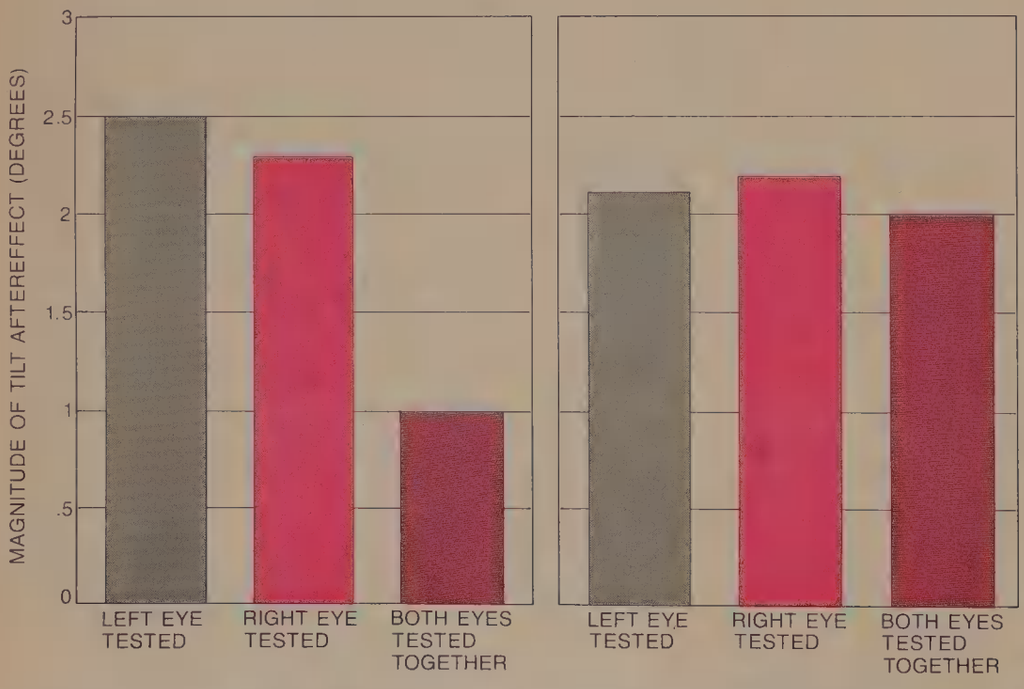
TILT AFTEREFFECT is an illusion that serves the study of hidden visual processes on the pathways leading to visual perception. To experience the illusion let your eyes scan back and forth along the horizontal black bar crossing the chevron at the left side of the illustration. You should then find that the two halves of the pattern at the right are briefly not colinear. The tilt aftereffect is measured by making the pattern at the right adjustable and asking subjects to adjust it so that it seems to be colinear immediately after they have stared at the chevron.

system had only the purely binocular process, you would be unable to see if one eye was closed. You would be unable to see when you tried, for example, to peek out from behind a tree and one eye surveyed only tree bark while the other surveyed a house.

The ability to perceive a purely binocular stimulus such as the cyclopean stimulus that served our study of vection does not ensure that the brain has a purely binocular visual process. The cyclopean stimulus could have been "seen" by a visual process that accepts input from both eyes and also from either eye alone. How, then, is it known that a purely binocular process is there? Again one resorts to an indirect method, in this case based on a temporary distortion of vision called the tilt aftereffect.

Look at the pattern of stripes at the right in the illustration above. The top half of the pattern should appear to be aligned with the bottom half. Now look at the chevron pattern at the left in the illustration. Let your eyes scan back and forth for one or two minutes along the horizontal black bar that crosses the center of the chevron, then quickly shift your gaze to the pattern at the right. You should find that the two halves no longer seem aligned. They should appear to be bent in the direction opposite to the direction of the bend in the chevron.

This is the tilt aftereffect. It can be measured by making the two halves of the pattern at the right adjustable and asking people to make the two halves look aligned. Before people view the chevron they make settings quite close to colinearity. After viewing the chevron their settings are systematically displaced by about two degrees.



PURELY BINOCULAR PROCESS is revealed by comparing the results of two experiments that employ the tilt aftereffect. In one of the experiments (*left*) subjects stared at the chevron with each eye open in alternation; thus the stimulus was available to every visual process that gets input from one eye and every process that gets input from either eye. On the other hand, the stimulus was unavailable to a purely binocular process: one that gets input only from both eyes. The subjects then manipulated the adjustable pattern with one eye or with both eyes open. The binocular aftereffect proved to be less than the monocular aftereffect. The second experiment (*right*) differed only in that the subjects stared at the chevron with both eyes open. The binocular aftereffect now proved to equal the monocular aftereffect. Evidently the binocular viewing of the stimulus in the second experiment exposed the purely binocular process.

Suppose a subject looks at the chevron with only his right eye open and then manipulates the adjustable pattern with only his left eye open. The left eye never sees the chevron, yet an aftereffect is detected. In short, there is interocular transfer. It is evidence for a binocular process but not a purely binocular one. After all, the process was activated when the right eye was exposed to the chevron and it was activated when the left eye was tested with the adjustable pattern. Apparently it can respond to either the left eye or the right.

As it happens, the aftereffect is smaller when it is tested with the eye that was not exposed to the chevron than it is when it is tested with the eye that was exposed. Only 70 to 80 percent of the aftereffect transfers. Randolph Blake and his colleagues at Northwestern University conclude from this finding that at least two processes are involved when the tilt aftereffect is tested with one eye. One of them is the binocular process; the other one is monocular. Thus the exposure of the right eye to the chevron activates both the binocular process and the right eye's monocular process. The testing of the right eye activates both of them again. The result is a strong tilt aftereffect. Suppose the left eye is tested. The binocular process, which was exposed to the stimulus, becomes active again. The left eye's monocular process also becomes active, but it never "saw" the chevron. Its output dilutes the magnitude of the aftereffect.

By means, then, of the tilt aftereffect Blake demonstrated the existence of both monocular and binocular processes. Held and I exploited the aftereffect to show that a purely binocular process exists as well. We had subjects look at the chevron with each eye in alternation: one minute with the left eye, one minute with the right eye, one minute with the left eye again and one minute with the right eye again. When we measured the aftereffect, we had them keep either the left eye open, the right eye open or both eyes open.

We reasoned that the alternating monocular viewing of the stimulus would expose every visual process that gets input from the left eye, every proc-

ess that gets input from the right eye and every process that gets input from either eye. If these were the only visual processes, they would constitute the entire visual system. Hence it would make no difference whether the subject had one eye open or both eyes open when he manipulated the adjustable pattern; the tilt aftereffect would be the same. If, however, there were a visual process that is active only when the left eye and the right eye are stimulated simultaneously, the results would be different. The alternating monocular viewing of the stimulus would leave this process unexposed, but then if both eyes were opened for the test of the aftereffect, the process would be activated, and it would dilute the aftereffect.

That is in fact what we found. The magnitude of the aftereffect was much less with both eyes open than it was with one eye closed. In a further experiment we had subjects look at the chevron with both eyes open. We expected this would expose every visual process, and we were right: the aftereffect was now much the same whether the subjects had both eyes open or one eye open when they manipulated the adjustable pattern.

In both experiments the left eye and the right eye were both exposed to the chevron. The only difference is that in the first experiment our subjects never saw the chevron with both eyes open at the same time. In the second experiment they did. A process was left unexposed by the first experiment and then was exposed in the second. That process must "see" only with both eyes. It is a purely binocular process.

What could a purely binocular process do in assistance of vision? In order to perceive three-dimensional depth the visual system seeks matches and slight mismatches between features in the images in each eye. The purely binocular process could serve in such a search. Visual perception cannot, however, be based on the output of a purely binocular process alone. If it were, all the unmatched parts of the two retinal images would vanish. In fact what we cannot see by means of a purely binocular process we can see by means of a monocular process or a process that responds to either eye.

From ancient times people have spoken of five senses. It is becoming ever clearer, however, that five is much too small a number. Senses such as touch seem to be divided into a variety of submodalities; the visual system is also divisible. Perhaps vision is best regarded not as a single sense but as a set of systems, each one a sense in its own right. It is likely that the full range of human visual senses remains to be discovered. The senses are like an old and crowded attic in that one finds the unexpected in each new corner one explores.

If you'd like to know more about Jack Daniel's Whiskey, drop us a line.

THESE MEN can tell exactly what's happening inside every barrel in a Jack Daniel's warehouse.

In the heat of summer the whiskey is expanding into the charred inner wood of the barrel. Come Halloween, it's starting to cool. And inching its way back toward the center. Over the aging period, this gentle circulation of whiskey is going on constantly. Of course, it can't be perceived by the human eye. But after a sip of Jack Daniel's, we believe you'll recognize its importance.

CHARCOAL MELLOWED
DROP
BY DROP

Tennessee Whiskey • 90 Proof • Distilled and Bottled by Jack Daniel Distillery
Lem Motlow, Prop., Inc., Lynchburg (Pop. 361), Tennessee 37352
Placed in the National Register of Historic Places by the United States Government.

The Active Solar Corona

Once visible only during eclipses, the pearly surround of the sun is under constant surveillance. The observations show dynamic activity reflecting the interaction of matter and magnetic fields

by Richard Wolfson

One of nature's rarest and most dramatic spectacles is the solar corona: the thin outer atmosphere of the sun that becomes visible during a total eclipse. Although a total eclipse of the sun occurs somewhere on the earth every year or so, most people have the chance to witness only one or two in a lifetime. Light from the solar corona is normally overwhelmed by light from the sun's luminous surface: the photosphere. That the corona is visible at all is a result of the remarkable coincidence that the moon's apparent size is almost exactly the same as the sun's. If the moon were any smaller, eclipses would never be total. If the moon were any larger, the intricate inner corona would be hidden from view.

In the past decade what is known about the corona and indeed about the sun itself has been vastly enlarged by instruments both on the ground and on spacecraft, unmanned and manned. The corona is observed to be in a continuous state of flux, punctuated by periodic explosive outbursts. A "wind" of coronal particles streams outward through the solar system; fluctuations in the wind reflect coronal activity. This view of the corona and its manifestations transforms the sun from a compact object at the center of the solar system to a single entity of giant dimensions, reaching beyond the most distant planets. The corona is a huge and yet nearby laboratory for studying astrophysical processes. Understanding the corona can help to advance knowledge of such diverse phenomena as fusion reactors here on the earth and supernova explosions in distant stars.

In ordinary photographs the corona appears as a pearly halo surrounding the eclipsed sun. Most of the light from the halo does not originate in the corona but is sunlight scattered from coronal electrons. For every million photons of light leaving the sun's photosphere only one is scattered in the corona. As a result the corona is a million times dimmer than the photosphere. The corona's transparency indicates that the coronal gas is exceedingly rarefied. Indeed, by laboratory standards on the earth the corona is an excellent vacuum; the maximum density of the coronal gas is between 10^8 and 10^9 particles per cubic centimeter.

Spectroscopic studies suggest that the corona is like the visible sun in composition, being dominated by hydrogen and helium, with only traces of the heavier elements. Such studies also offer evidence of the corona's temperature. When gases are heated, their atoms become ionized, or shorn of one or more electrons. The coronal gas turns out to be much more highly ionized than the gas in the photosphere, whose temperature is some 6,000 degrees Kelvin (degrees Celsius above absolute zero). In the corona hydrogen and helium are essentially stripped of electrons, and the heavier elements have lost many of their electrons. The high state of ionization signifies that the corona is extremely hot. Moreover, the lines of the coronal spectrum are broadened, indicating that coronal particles are in random motion at high speed. On the basis of these spectroscopic observations the temperature of the corona is estimated to lie between one and two million degrees K.

Accounting for the high temperature of the corona long presented theorists with a major puzzle. The second law of thermodynamics states that heat cannot flow spontaneously from a cooler object to a hotter one. For this reason the thermal energy of the sun, enormous though it is, cannot heat the corona beyond 6,000 degrees K. So fundamental is this objection that for decades solar physicists were reluctant to accept the spectroscopic evidence at face value. Various ad hoc hypotheses were put forward, suggesting, for example, that the ions in the corona might originate somehow deep within the sun or that they might be created directly in the corona by nuclear fission. Finally in the 1940's the concept of a truly hot corona became widely accepted. Once theorists were convinced of the reality of high coronal temperatures they soon made rapid progress in accounting for the corona's gross physical structure.

If the second law of thermodynamics rules out the transfer of heat energy from the photosphere to the corona, what is the heating mechanism? The most likely candidate is the direct transfer of mechanical energy. According to present theories, mechanical energy that originates in the convective layer of the sun below the photosphere is carried outward to the corona in the form of either intense sound waves or magnetohydrodynamic waves.

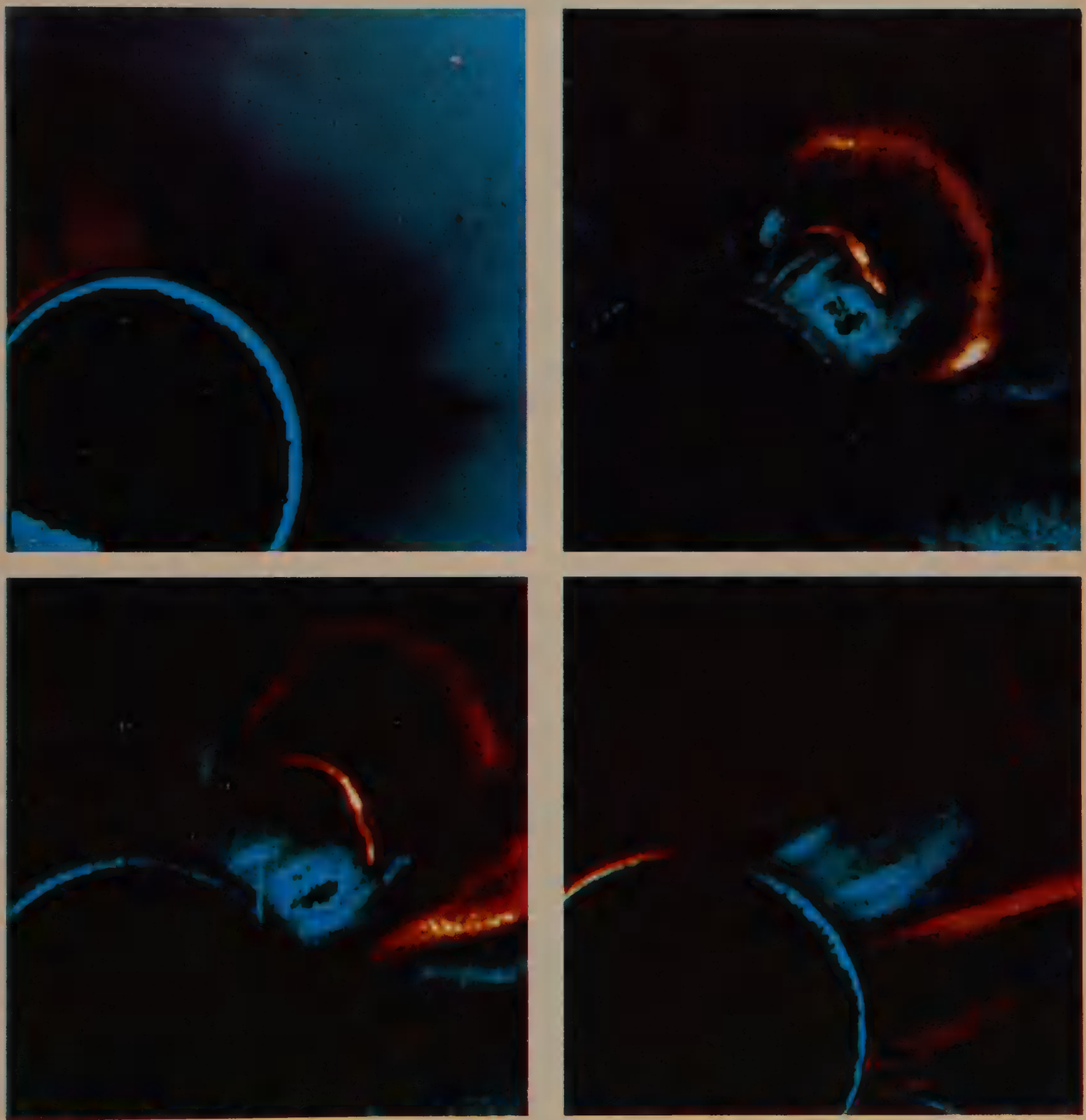
Sound waves are simply mechanical disturbances of a gas that manifest themselves as pressure variations. Magnetohydrodynamic waves, which occur in ionized gases, are disturbances arising from the interaction of ionized matter and magnetic fields. Although the general concept of mechanical heating by waves is well established, the elaboration of detailed models consistent with observation is engaging the efforts of many investigators. The work promises to link the physical properties of the corona with those of the underlying layers, and it has the goal of achieving a unified understanding of the entire sun.

The corona, then, is a region of high temperature and low density that surrounds the cooler, denser and more familiar photosphere. (A thin intermediate layer, the chromosphere, separates the photosphere from the corona.) Most ordinary photographs show that the corona is far from symmetrical, often appearing dimmer and less extended near the solar poles. Away from the poles prominent bulges suggest a complexity of structure. Such deviations from symmetry vary from eclipse to eclipse. The most symmetrical coronas coincide with times of maximum sunspot activity in the 11-year solar cycle.

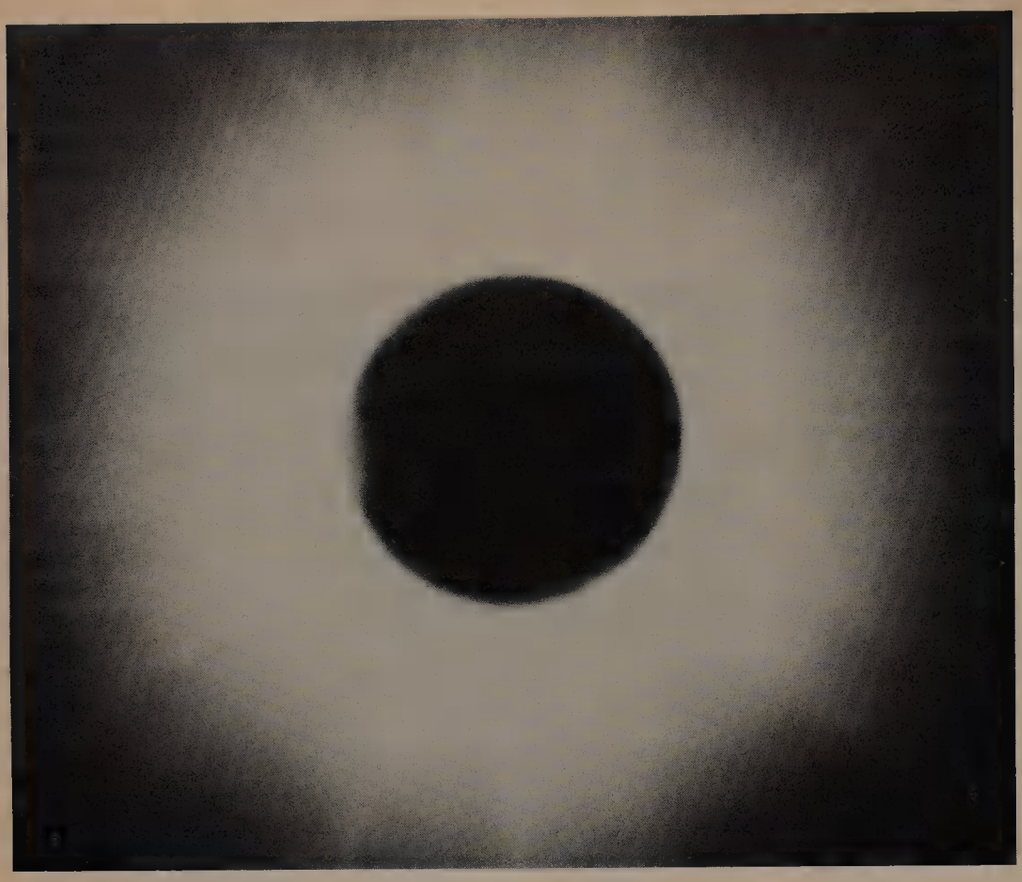
Coronal structure hinted at in ordinary photographs is largely washed out by overexposure of the bright inner corona. The most successful method for overcoming this limitation is to photograph the corona through a special fil-

ter whose density is radially graded to block more light at the center of the image than at the edge, thereby compensating for the falloff in coronal brightness. With such filters coronal structures can be detected at distances of many solar radii from the photosphere.

In a photograph taken through a radially graded filter the delicate beauty of the solar corona is revealed in intricate detail. Long, graceful streamers extend outward for several million kilometers. Closer to the solar disk are bright loops of coronal material. In some pictures straight, narrow rays can also be seen. At times of maximum solar activity such coronal features surround the entire solar disk. At times of minimum activity, in contrast, coronal features are confined to the lower solar latitudes. At such times the corona shows large



TRANSIENT LOOP IN THE SOLAR CORONA was recorded in April, 1980, by the coronagraph aboard the Solar Maximum Mission (SMM) satellite. A coronagraph incorporates a disk that blocks the brilliant light from the sun's photosphere, or visible surface, so that the million-times-dimmer corona can be observed. The coronagraph carried by the SMM spacecraft was developed by the High Altitude Observatory of the National Center for Atmospheric Research in Boulder, Colo. The light of the corona is focused on a television camera tube, and the image is transmitted back to the earth in digital form. In this four-part sequence, displayed in arbitrary colors, the corona before the transient event appears at the upper left. The other three pictures are made by a digital subtraction technique in which the light value at each point represents the change in light intensity from the level recorded in the first image. At the upper right the coronal loop has attained nearly the diameter of the sun. At the lower left, 25 minutes later, the loop, expanding at the rate of about 300 kilometers per second, has reached the edge of the frame, and behind it a solar prominence has begun to move upward through the corona. At the lower right, 20 minutes later, only one leg of the loop is visible. It is not known whether the eruptions of the coronal loop and the solar prominence are causally related or whether both of them are triggered by some underlying instability in the magnetic field of the sun.



IN A CONVENTIONAL PHOTOGRAPH OF A SOLAR ECLIPSE the inner corona is overexposed and all delicate detail is washed out. The photograph was made in India on February 16, 1980, by a group from Williams College led by Jay M. Pasachoff. The missing detail is revealed in the photograph below. The conventional photograph nonetheless shows that the corona is symmetrical, a feature associated with maximum activity in the 11-year sunspot cycle.

DELICATE FEATURES OF THE CORONA during the eclipse of February 16, 1980, were captured by interposing a radially graded filter between the telescope lens and the photographic plate. By sharply reducing the brilliant light from the inner corona the filter makes it possible to record coronal features from the photosphere out to a distance of several solar radii. The bright spots just above the disk of the moon are prominences, consisting of cool, dense material supported in the corona by magnetic fields. Many well-defined coronal streamers are visible. The photograph was made in India by a group from the High Altitude Observatory.

"holes" at both poles, where the corona is much dimmer than it is near the equator. Within the polar holes faint polar "plumes" are often visible. A bundle of these narrow structures diverges from each pole, fading to invisibility within one solar radius or so. The dominant features—streamers, loops and holes—are so prevalent that some investigators believe the corona consists only of such features and there is no such thing as an underlying "quiet corona."

The appearance of the corona varies significantly from one eclipse to another. Even ordinary eclipse photographs are sufficiently different for an observer usually to be able to identify a specific eclipse from its photograph. Coronal images made through radially graded filters vary dramatically. In addition to the general changes associated with the sunspot cycle, each eclipse has its own particular configuration of coronal streamers, loops and other features. A set of photographs made at different times during a single eclipse provides tantalizing evidence for a dynamic corona, whose features change on a time scale much shorter than the interval between eclipses.

In order to study coronal dynamics one would like to observe the corona even when there is no eclipse. Given the brightness of the solar photosphere, this would seem a hopeless task. At visible wavelengths the dense photosphere is a million times brighter than the tenuous corona. Since the corona is much hotter than the photosphere, however, it produces high-energy forms of electromagnetic radiation that are virtually absent in the spectrum of the photosphere. At two million degrees the coronal gas is a strong source of X rays. Except for the X rays released sporadically by solar flares and other local "hot spots" the cooler photosphere is a negligible source of such rays. Accordingly the corona can be observed by pointing an X-ray telescope directly at the sun.

Such observations must be made from space, because the earth's atmosphere is opaque to X rays. The first X-ray telescopes were flown on brief rocket flights in the early 1970's and made crude images of the corona that showed bright areas suggesting regions of high density or temperature. By comparing eclipse photographs made from the ground with X-ray images made from space at about the same time the correspondence between X-ray-bright regions and optically visible coronal features was confirmed.

X-ray imaging technology has advanced rapidly over the past decade. In the mid-1970's astronauts aboard the Skylab orbiting observatory obtained many X-ray images of the corona. A motion picture made from such images shows clearly how coronal features

Wir sprechen die Sprache derer, die unsere Zukunft prägen.

The German-language edition of SCIENTIFIC AMERICAN, SPEKTRUM DER WISSENSCHAFT, has a fully paid circulation that has surged in only two years from 60,000 to 92,000* today.

German industry has evolved as a formidable competitor to leading American, Japanese and other European companies because its technology is of the highest caliber. The well-educated, affluent professionals that read SPEKTRUM DER WISSENSCHAFT are making the decisions and shaping the opinions of German businesses through their technical expertise.

In business, industry, government and public life, SPEKTRUM's readers are emerging as the most influential members of their nation's professional-managerial population.

We speak the language of more than 3,000,000 world shakers in 140 countries each month. We reach the people who make the future happen in German, French, Spanish, Italian, Japanese, Chinese and English.

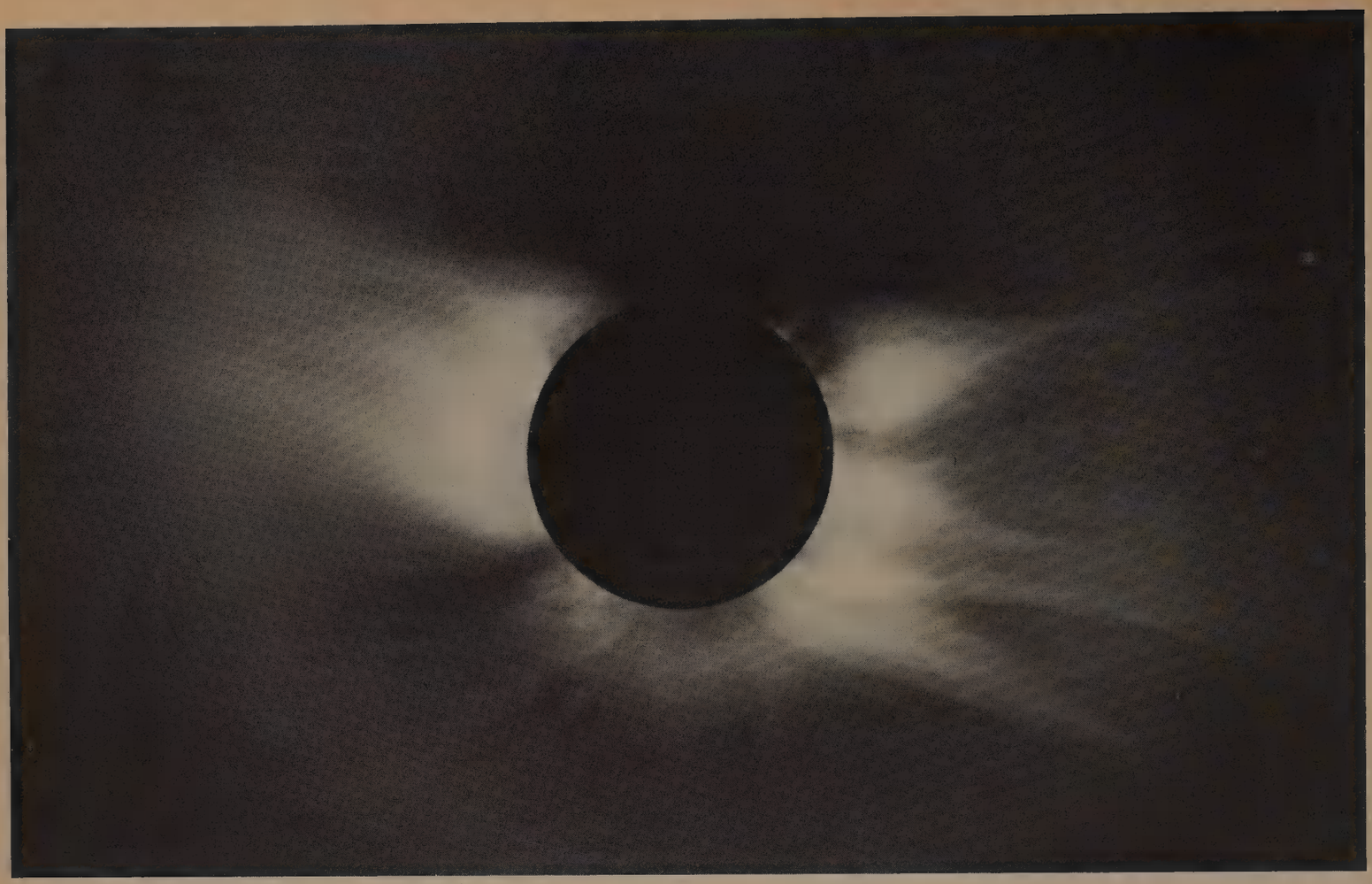
For more information contact:

SPEKTRUM DER WISSENSCHAFT
Anzeigendirektion Deutschland
Mönchhofstr. 15
D-6900 Heidelberg
West Germany
Telephone (011-49) 6221-46037/38

*IVW audited

SCIENTIFIC AMERICAN

We reach the people who make the future happen.



HIGHLY ASYMMETRIC CORONA was photographed through a radially graded filter during the eclipse of June 30, 1973, when the 11-year sunspot cycle was near a minimum. At such times coronal streamers are concentrated near the solar equator, and the solar poles exhibit "holes": regions of depressed activity. The photograph was made in Kenya by workers from the High Altitude Observatory.

evolve and move with the sun's rotation. Skylab X-ray images were instrumental in establishing that coronal holes exist not only at the poles but also on a smaller scale throughout the corona. Improved X-ray-imaging and spectroscopic instruments on recent spacecraft have supplied data of still higher resolution on the X-ray corona.

Additional information about the corona can be gleaned from the ultraviolet radiation emitted by highly ionized heavy elements in the corona. At these wavelengths too the output of the photosphere is negligible. Ultraviolet data are valuable because particular spectral lines are formed only within narrow temperature ranges, making it possible to probe different layers of the corona, the chromosphere and the transition region between them.

At the other end of the spectrum ground-based radio telescopes can identify and track signals associated with characteristic oscillations of the ionized coronal gas. The frequency of such oscillations is a direct measure of the density of the corona, so that different coronal regions can be probed simply by changing the frequency to which the radio telescope is tuned. By linking several telescopes into an interferometer the evolution of individual coronal features can be followed. At Culgoora in Australia the Commonwealth Scientific and Industrial Research Organization (CSIRO) has set up an instrument called a radioheliograph, consisting of 96 interconnected radio telescopes dedicated to solar studies.

The most direct approach to investigating the corona in the absence of an eclipse is to contrive an artificial eclipse. In principle a thumb held at arm's length could block the light of the photosphere. The reason this does not work is that the earth's atmosphere scatters enough sunlight to overwhelm the coronal light. Above the atmosphere, however, and even on high mountains it is possible to observe the corona by artificially blocking the photosphere. An instrument designed for this purpose is the coronagraph. It basically consists of a small refracting telescope equipped with a disk that occults the sun as the moon does in an eclipse. It includes, however, a number of other occulting devices designed to minimize scattered light and diffraction effects that distort the coronal image. Once the coronal light is isolated it can either be recorded directly or passed through polarizers or filters to isolate special features.

The earliest coronagraphs were developed some 50 years ago. Since then they have been installed at mountaintop observatories, flown on balloons and launched into space. A coronagraph on Skylab discovered coronal transients: rapidly evolving features that appear to eject coronal material into space. There are currently coronagraphs aboard several unmanned spacecraft. The major ground-based coronagraphs are at the Sacramento Peak Observatory in New Mexico, on the Pic du Midi in France and on Mauna Loa in the Hawaiian Islands. In 1979 an orbiting coronagraph operated by the Naval Research Laboratory recorded an event never observed before: a comet colliding with the sun. The newest coronagraphs record their images by means of digital electronic detectors rather than on photographic film. Data from the electronic detectors can be processed by computer to yield enhanced images with high resolution and contrast.

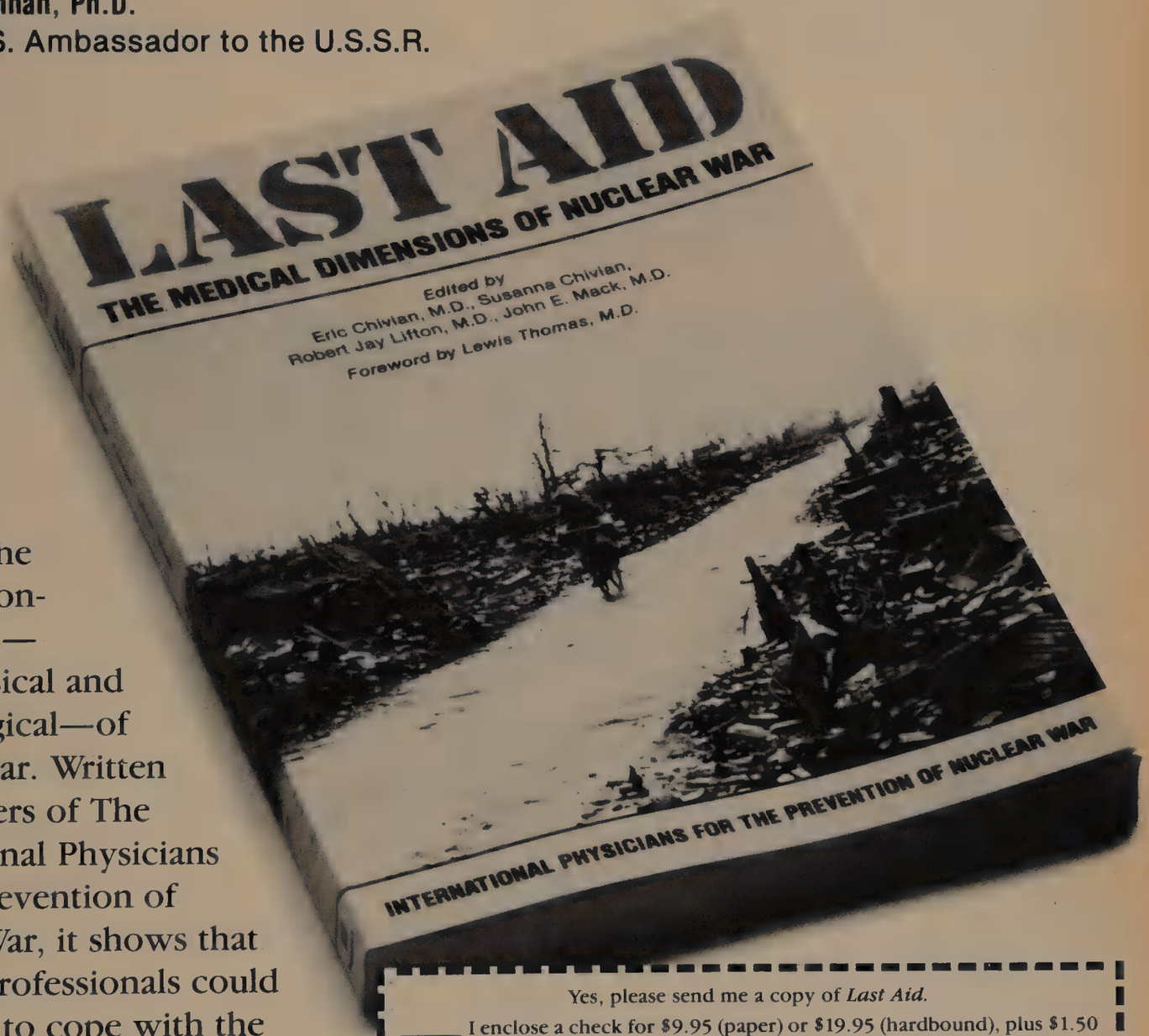
Some of the best coronal images come from a coronagraph on the Solar Maximum Mission satellite. The SMM was launched in February, 1980, near the peak of solar activity in the current sunspot cycle. The spacecraft carries an array of instruments designed for coordinated study of the sun at wavelengths from visible light to gamma rays. For a

"To ignore this passionate presentation is to turn our backs on the most urgent issue of our time."

Walter Cronkite

"The facts revealed and documented in this book are fundamental for any constructive thinking about the problems of nuclear weaponry, and deserve the widest possible attention."

George F. Kennan, Ph.D.
Former U.S. Ambassador to the U.S.S.R.



Last Aid presents the medical consequences—both physical and psychological—of nuclear war. Written by members of The International Physicians for the Prevention of Nuclear War, it shows that medical professionals could not begin to cope with the overwhelming problems that would ensue in the aftermath of a nuclear holocaust. A reaction to heads of government who speak rationally of a conflict that can have no rational basis, ***Last Aid*** **is must reading for every intelligent, thinking human being.**

Yes, please send me a copy of *Last Aid.*

_____ I enclose a check for $9.95 (paper) or $19.95 (hardbound), plus $1.50 to cover postage and handling.
(California residents add appropriate sales tax.)

_____ Please charge my ☐ MasterCard ☐ VISA

Account # _____ Expires _____

Signature _____
(All credit card orders must be signed.)

Name _____

Address _____

City _____ State _____ Zip _____

W. H. Freeman and Company
660 Market Street
San Francisco, CA 94104

ISBN hardbound 0-7167-1434-5
ISBN paper 0-7167-1435-3
SA

little more than six months the SMM yielded a wealth of data, but then a series of failures in the system that points the spacecraft at the sun left it spinning out of control, cutting short a mission that was projected to last for two years. The SMM, however, is the first spacecraft designed to be repaired in space. It is equipped with a grappling hook and its instrument packages are replaceable. The Space Shuttle mission scheduled for April, 1984, will include a rendezvous with the SMM during which an astronaut will spacewalk to the ailing satellite and stop it from spinning. Once the SMM is stabilized the shuttle's mechanical arm will bring it into the shuttle's cargo bay, where it will be repaired and then redeployed in its orbit. The revived SMM should resume its observations during the sunspot minimum, providing valuable data to compare with those gathered during the maximum.

The coronagraph on the SMM was developed by a group under the direction of Lewis L. House at the High Altitude Observatory of the National Center for Atmospheric Research. In it coronal light is passed through interchangeable filters and polarizers and focused on a television camera tube. Data from the camera are transmitted to the ground in digital form and stored on magnetic tape. Each image from the instrument includes well over a million bits of information and shows one quadrant of the corona. On the ground the images are processed by computer and displayed on a color television monitor. The computer enables the observer to zoom in on selected areas of the image and to enhance particular features.

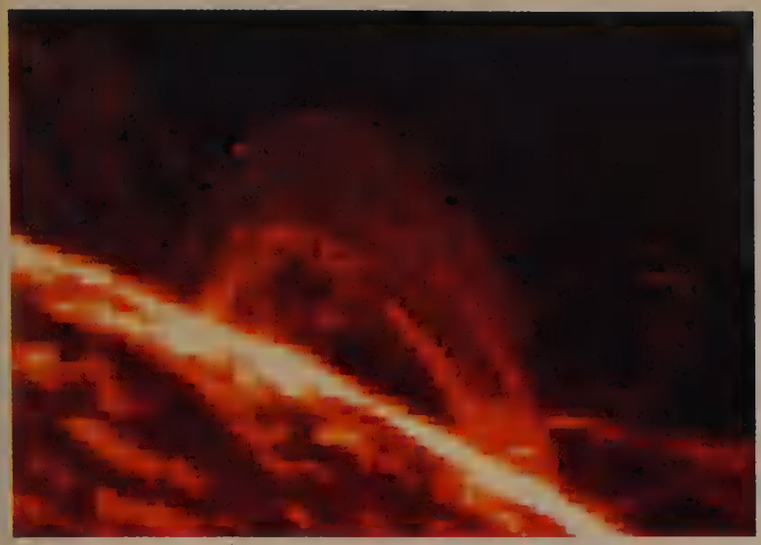
Before its temporary demise in the fall of 1981 the SMM coronagraph sent back some 30,000 coronal images. A motion picture prepared by Rainer M. E. Illing of the High Altitude Observatory from a sequence of images taken in April, 1980, reveals the corona as it has not been seen before: a restless, dynamic structure of ever changing aspect. Although major coronal features remain recognizable for several days or longer, they clearly evolve from day to day. Smaller features come and go on shorter time scales. Occasional transient events disrupt the corona in a matter of hours. A typical transient event captured by the SMM coronagraph shows a coronal loop suddenly expanding and arching outward at a rate of hundreds of kilometers per second.

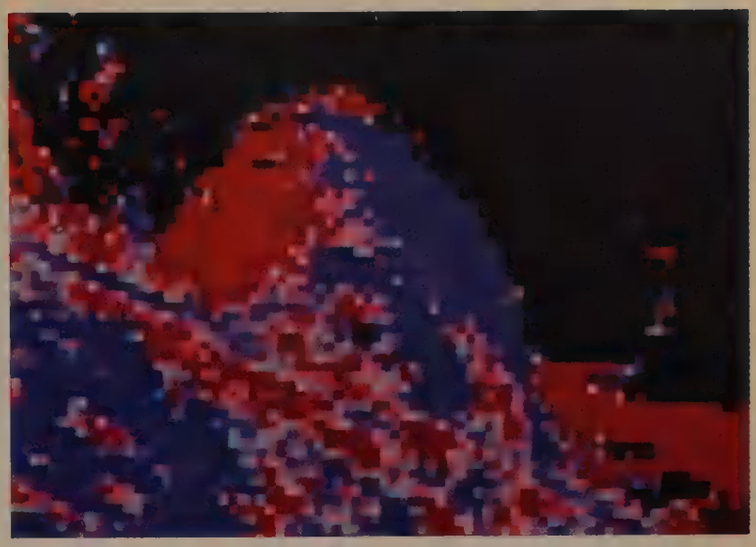
X-RAY IMAGE OF THE SUN records radiation that originates predominantly in the corona, where the temperature is about two million degrees Kelvin (degrees Celsius above absolute zero). The sun's photosphere, at 6,000 degrees K., is much too cool to emit X rays. The X-ray image was made by astronauts aboard Skylab on September 18, 1973, about two and a half months after the solar eclipse shown in the illustration on page 108. Most bright areas in the X-ray image are regions of enhanced coronal density or temperature, corresponding to coronal streamers or loops. The dark areas are coronal holes. In addition to the polar holes a prominent coronal hole can be seen near the equator. Images such as this one demonstrated that coronal holes are not limited to the polar regions of the sun and that they are associated with high-speed streams in the wind of particles from the sun. The image was provided by American Science & Engineering, Inc., which designed the X-ray telescope aboard Skylab.

What is the physical nature of coronal features? How can a hot, tenuous gas organize itself into such distinct patterns? Why do some coronal features persist for days and others evolve suddenly and rapidly? Neither the spherically symmetrical force of gravity nor the forces associated with the rotation of the sun can mold the corona into its intricate forms. A clue to the origin of coronal structures is given by the forms themselves. Individual coronal loops and polar plumes suggest the patterns formed when iron filings are sprinkled near a magnet. Just as the filings align themselves with the lines of magnetic force, so the structure of the sun's magnetic field imposes itself on the coronal gas. Magnetic force is responsible for the detail, variety and evolution of coronal structure.

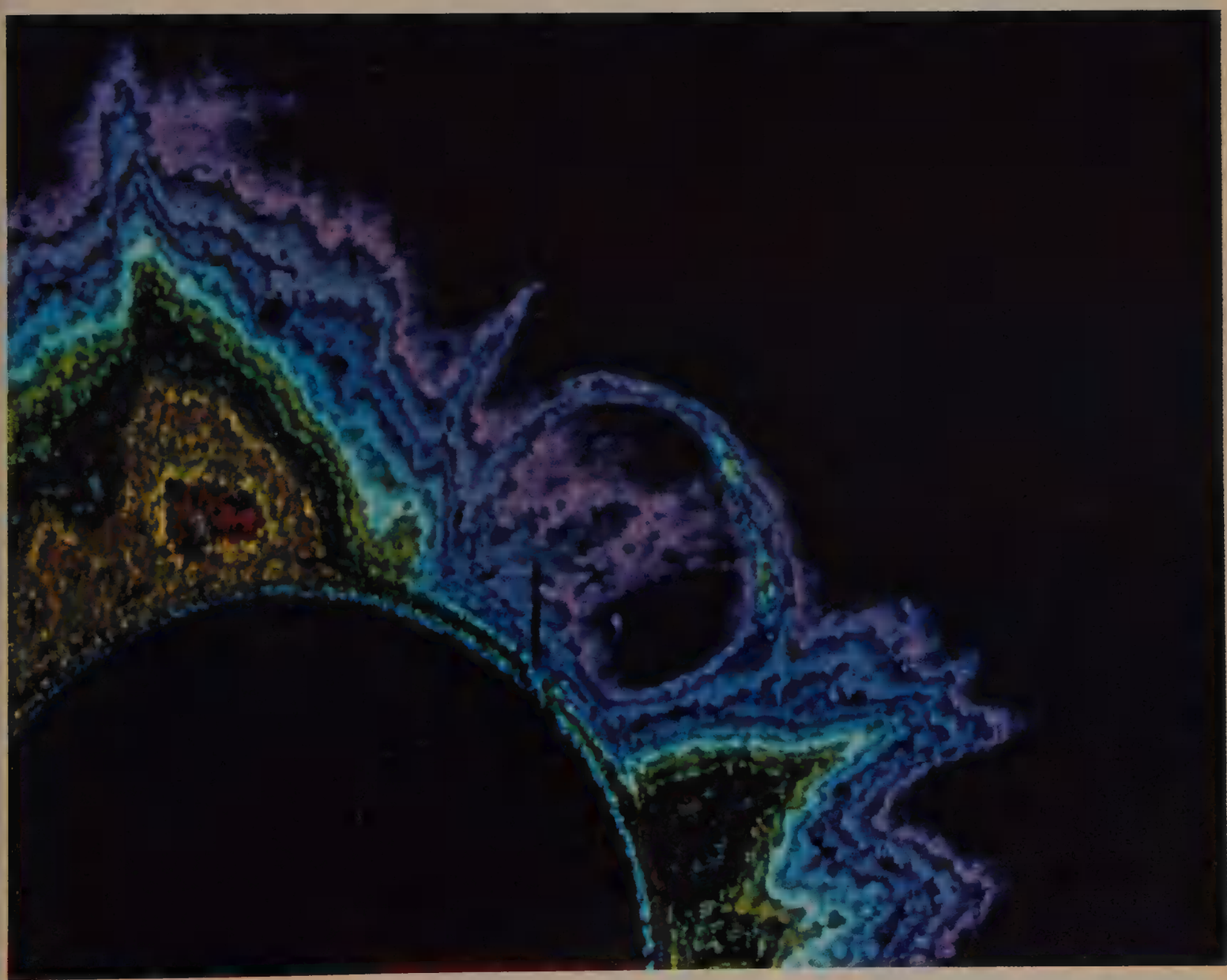
Like the magnetic field of the earth and many other celestial objects, the sun's magnetic field is thought to originate in the interaction of convective motion and rotation of the electrically conducting material in the object's interior. For reasons that are still poorly understood the sun's field is highly variable. Roughly every 11 years the sun's north and south poles reverse. The magnetic poles are strongest near sunspot minimum and give the solar magnetic field the simple overall structure of a magnetic dipole such as a bar magnet or the earth. The dipole structure is evident in the polar plumes seen in the corona near sunspot minimum. At sunspot maximum the solar magnetic field is chaotic, showing no distinct poles but rather many local regions of intense magnetism, often associated with the sunspots. The coronal loops, which at this time are particularly numerous, outline regions of strong local magnetic fields.

The interaction of the coronal gas and the magnetic field is governed by the basic laws of electromagnetism. We are most familiar with magnetism through the behavior of ferrous metals such as iron. Fundamentally, however, a magnetic field is generated by moving electric charges. In magnetic materials the charges are electrons moving in orbit around atomic nuclei, collectively generating the magnetism of the entire piece of material. In an ionized gas such as the corona the constituent particles of the gas are charged and hence interact with magnetic fields. These electromagnetic interactions make ionized gases so

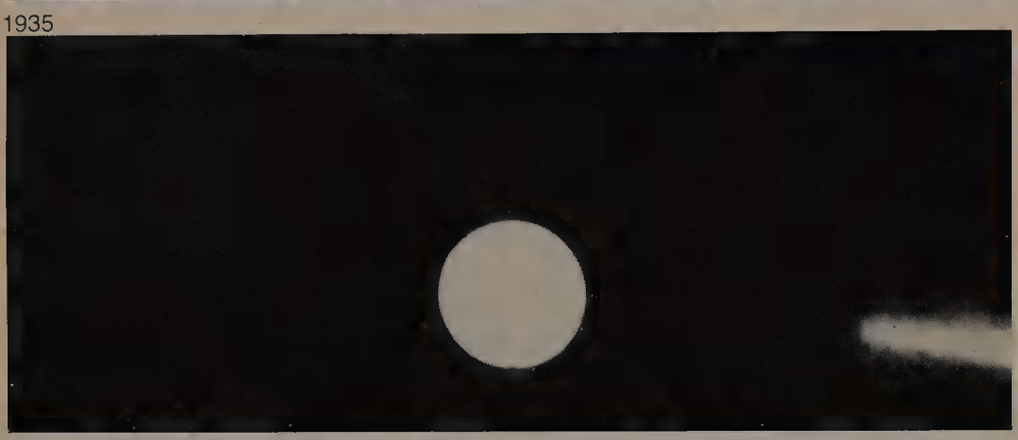


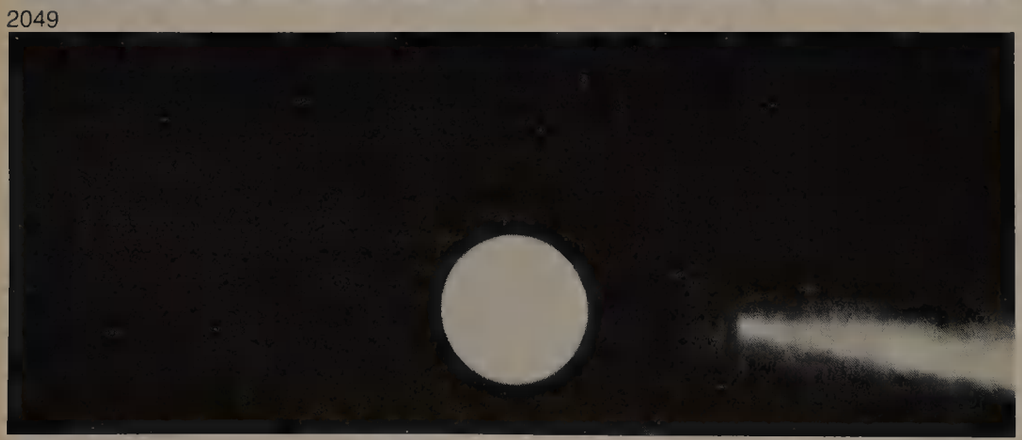


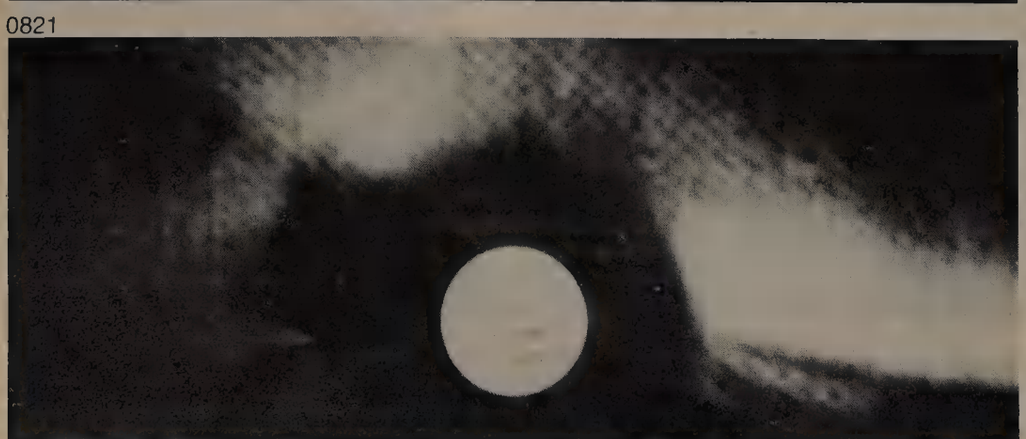
LOOP NEAR THE PHOTOSPHERE, supported by a strong local magnetic field, has a temperature of about 100,000 degrees K. The images record the ultraviolet radiation at a wavelength of 1,548.19 angstrom units emitted by triply ionized carbon: atoms of carbon stripped of three electrons of their normal complement of six. The image at the left is formed from all the radiation in a narrow band centered at that wavelength. The image at the right is a "velocitygram" of the same loop made by splitting the radiation into two component wavelengths, one slightly shorter than the central wavelength and one slightly longer. The shorter component, represented in blue, is from matter approaching the observing instrument (the ultraviolet spectrophotometer on the Solar Maximum Mission satellite); the longer component, in red, is from matter receding. A possible interpretation of the image is that the right half of the loop is the closer of the two and that material is flowing down toward both feet. Such velocitygrams reveal the dynamic character of the sun's outer layers.



CORONAL STREAMERS and a rising prominence loop were recorded by the High Altitude Observatory coronagraph aboard the Solar Maximum Mission satellite on April 14, 1980. The contrast of the image was enhanced by "effective contouring," one of several computer techniques for bringing out details. The different colors in the image correspond to different values of the density of the corona. This image is actually a high-contrast version of the coronal event appearing in the four-part sequence reproduced on page 105.







different from ordinary gases that an ionized gas has acquired a special name: plasma. The plasma state is often called the fourth state of matter. On our relatively cool planet earth plasmas are rare. Elsewhere in the universe, however, most matter is believed to exist in plasma form.

A charged particle in a magnetic field is subjected to a force that depends on its charge, its velocity and the strength and direction of the field. The direction of the magnetic force is at right angles to both the direction of motion of the particle and the direction of the field. The force is greatest when the direction of motion and the direction of the field are themselves at right angles, and it diminishes to zero when the direction of motion and the direction of the field are parallel. Since the force is always at right angles to the direction of the particle's motion, it can change only that direction and not the particle's speed. And since there is no force when the motion is parallel to the field, only those components of particle motion that are at right angles to the field are affected.

The net result is that charged particles move in helical paths around lines of force in the magnetic field. A particle is free to move in the direction of the field, but if it is pushed at right angles to the field, it moves in a circle rather than in a straight line. Charged particles in a magnetic field are said to be "frozen" to the field, much like beads on a wire. The particles are free to move only along the direction of the field. Nonuniform fields and collisions between particles make the freezing less than perfect, but in a thin plasma such as the corona the assumption that the plasma is frozen to the field is an excellent approximation.

The many charged particles in the corona constitute a gas that exerts pressure just as the earth's atmosphere does. Imagine the coronal gas trying to push its way under pressure into a region where there is a magnetic field. Since the plasma cannot move readily across the lines of force in the field, the field effectively counteracts the gas pressure. A more quantitative analysis shows that a magnetic field indeed exerts pressure on an ionized gas, with the magnetic pres-

COMET COLLIDING WITH SUN, the first recorded event of its kind, was pictured by an orbiting coronagraph operated by the Naval Research Laboratory. The sequence (*top*) begins at 1856 Universal Time on August 30, 1979. The sun itself is completely obscured by the disk of the coronagraph; the white disk represents the sun's size and location. The comet, which is one of the family known as sungrazers, was never detected by earth-based observers. The brightness after the "splashdown" persisted for some 24 hours. Since 1979 two more such collisions have been observed.

THE INPUT/OUTPUT STRUCTURE OF THE UNITED STATES ECONOMY

WHAT MAKES THE U.S. ECONOMY TICK?

The editors of SCIENTIFIC AMERICAN have prepared a wall chart displaying for the 1980's the Input/Output Structure of the U.S. Economy based on the latest interindustry study from the U.S. Department of Commerce.

The SCIENTIFIC AMERICAN Input/Output wall chart does for economics what the table of elements does for chemistry. It answers at a glance questions about the linkage between the microeconomics of the firm and the macroeconomics of the system; about the web of technological interdependencies that tie industry to industry; about the industry-by-industry direct and indirect consequences of swings in public and private spending; about the impact of change in technology, and about any other topic you can think of. You are rewarded by surprise as well as by confirmation of your hunches. For teaching and practical and theoretical studies, here is a powerful, graphic tool.

In the familiar format of the SCIENTIFIC AMERICAN Input/Output wall charts for the 1960's and 1970's, the wall chart for the 1980's measures 65″ × 52″ and is printed in eight colors. Each of the nearly 10,000 cells in the 97-sector interindustry matrix shows (1) the interindustry commodity flow, (2) the direct input/output coefficient and (3) the "inverse" coefficient. Where the direct input/output coefficient exceeds .01, the cell is tinted in the color code of the industrial bloc from which the input comes. This device, combined with triangulation of the matrix, brings the structure of interindustry transactions into graphic visibility.

A supplementary table displays, industry by industry, the capital stock employed; the employment of managerial, technical-professional, white-collar and blue-collar personnel; the energy consumption by major categories of fuel, and environmental stress measured by tons of pollutants.

The editors of SCIENTIFIC AMERICAN are happy to acknowledge the collaboration, in the preparation of this wall chart, of Wassily Leontief, originator of input/output analysis—for which contribution to the intellectual apparatus of economics he received the 1973 Nobel prize—and director of the Institute for Economic Analysis at New York University.

Packaged with the chart is an index showing the BEA and SIC code industries aggregated in each of the 97 sectors.

SCIENTIFIC AMERICAN
415 Madison Avenue, Dept. I/O, New York, N.Y. 10017

YES, send me the new SCIENTIFIC AMERICAN 1980's INPUT/OUTPUT wall chart—printed in eight colors on heavy paper measuring 65″ × 52″.

I understand my chart(s) will be shipped prepaid, uncreased*, flat-rolled and in a sturdy mailing tube. (Allow 4-6 weeks for delivery)

I enclose a check for $____ for ____ chart(s) at $35 each, saving me $5 each in handling charges.

Bill me for $____ for ____ chart(s) at $40 each.

Name_____

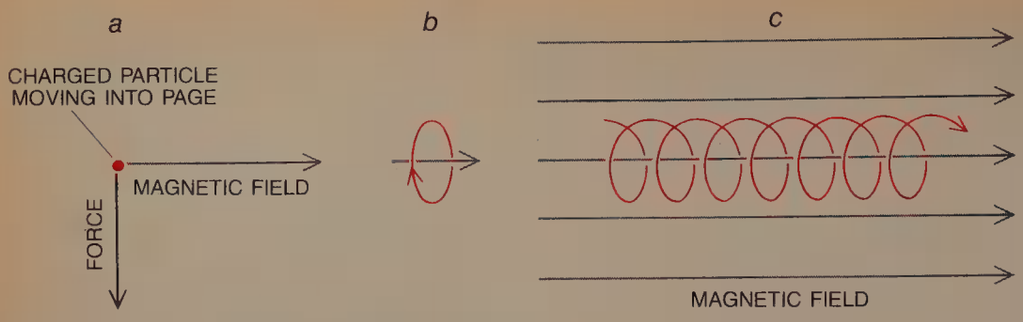
Company_____

Street Address_____

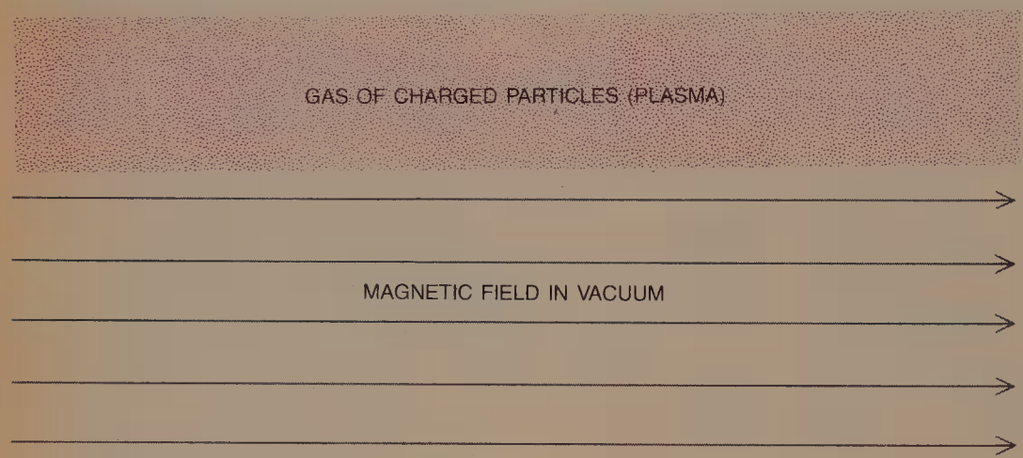
City, State, Zip Code_____

*Charts shipped international require one fold. 2/83

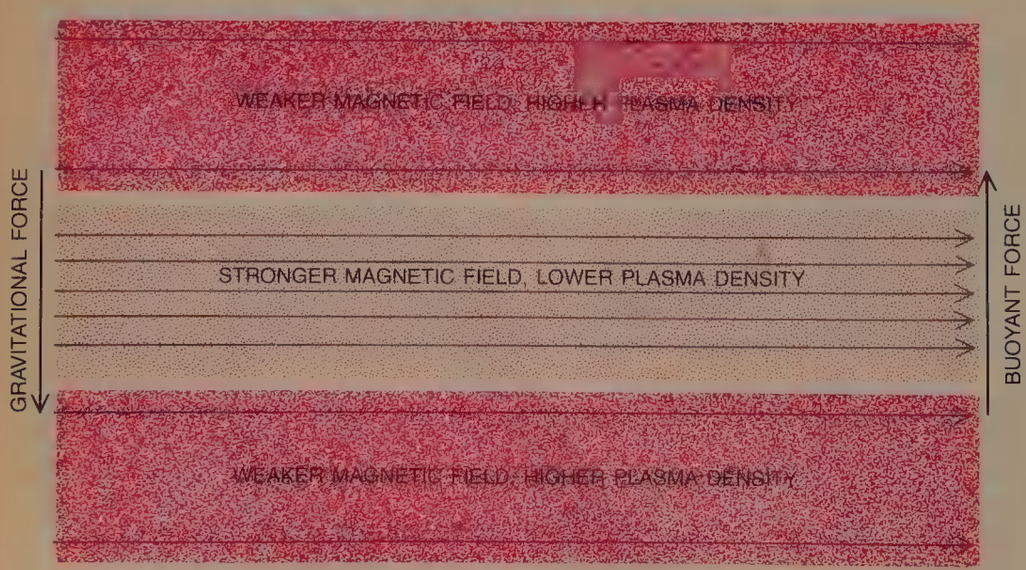
Make check payable to SCIENTIFIC AMERICAN. Please add sales tax for California, Illinois, Massachusetts, Michigan and New York.



CHARGED PARTICLES IN THE CORONA exhibit behavior characteristic of all charged particles under the influence of a magnetic field. The force on such a particle is at right angles to both the direction of the field and the direction of the particle's motion. For a positively charged particle traveling directly into this page (*a*) in the presence of a magnetic field pointing to the right the force on the particle would be directed downward. As the particle moves, the force continuously deflects it from a straight path so that it assumes a circular orbit (*b*). If the particle also has a component of velocity in the direction of the field, the particle's trajectory becomes a helix (*c*). The particle can move in either direction parallel to the field, but in response to a force acting at a right angle to the field, the particle moves to a helix of larger diameter. A particle of negative charge would circle in a direction opposite to the one depicted here.



MAGNETIC PRESSURE arises because plasma, a gas of charged particles, cannot move readily at right angles to a magnetic field. In the simple case shown here a vacuum containing a magnetic field is immediately adjacent to a plasma under pressure. The plasma is prevented from entering the vacuum by the pressure of the magnetic field. Although the plasma pressure and the magnetic pressure are assumed to be in balance, the configuration is not necessarily stable. A small disturbance in either the plasma or the field can trigger a drastic rearrangement.



MAGNETIC BUOYANCY is exhibited by a region of higher magnetic field strength that is sandwiched between regions of lower strength when all three regions are in a gravitational field. Because the total pressure in each region is the sum of magnetic and plasma pressures, the region of higher magnetic field strength must have a lower plasma pressure if it is to be in pressure equilibrium with its surroundings. If the regions are at the same temperature, lower pressure implies lower density. Since the less dense region is less tightly held by gravity, it tends to float upward. Such buoyancy may account for events in which loops are ejected into space.

sure being proportional to the square of the strength of the field. The behavior of the coronal gas is vastly more complicated than that of the earth's atmosphere, since both gas pressure and magnetic pressure determine the dynamics of the gas.

The coronal plasma does not merely respond passively to the solar magnetic field. Since it is highly ionized and tenuous, the corona is an excellent electrical conductor and can sustain large electric currents. The currents in turn give rise to magnetic fields that modify the original fields. This two-way interaction between the corona and the sun's magnetic field gives coronal structure a richness and variety that is simply unavailable to a nonconducting gas.

The richness of coronal phenomena greatly complicates attempts to explain coronal behavior theoretically. A full understanding of the corona calls for the self-consistent solution of a complicated set of equations in which the magnetic field shapes coronal structure even as coronal currents are modifying the field. In addition the effects of gravity, rotation and ordinary gas pressure cannot be neglected. Needless to say, a full theoretical understanding of all coronal features seen in a particular eclipse or coronagraph image has yet to be achieved.

An important first step has been taken, however, in a mathematical model developed by Gerald W. Pneuman and Roger A. Kopp of the High Altitude Observatory; the model clearly shows the development of coronal streamers at the solar equator. Refinements of equatorial-streamer models by Pneuman, Tyan Yeh of the National Oceanic and Atmospheric Administration and others have led to a detailed understanding of streamer structure. Current research in this area includes attempts to model the association of a coronal streamer and an adjacent coronal hole. A model I have been developing deals with the formation of streamers at high solar latitudes.

The discoveries of the past decade have stimulated much theoretical effort to understand the forces behind coronal dynamics, particularly very rapid phenomena such as coronal transient events. Recent work by B. C. Low of the High Altitude Observatory and me suggests that magnetic buoyancy may be responsible for the uplifting of coronal loops in transient events. Just as a piece of wood is subjected to buoyant forces when it is immersed in water, so regions of stronger magnetic field in the corona are buoyed up by the surrounding gas. Theoretical models show that a coronal loop can suddenly become unstable and rise through the corona either because of the random motion of photospheric gas at the base of the magnetic loop or in response to a violent photospheric event such as a solar flare.

Since the coronal plasma and magnet-

PONTIAC 2000

5-SPEED, OVERHEAD CAM, FUEL INJECTION

Here's a combination that could only mean one thing: pure driving fun!

Check out the way the front-wheel drive takes you through the turns. The way the reclining front bucket seats hold you close. And the way the new standard 5-speed manual transmission with overdrive takes you from first to fantastic... in a hurry.

To get you there with even more excitement, our sleek and sporty Pontiac 2000 is equipped with an overhead cam 4-cylinder engine with electronic fuel injection. And for responsive road handling, you get rack and pinion steering, MacPherson strut front suspension and semi-independent rear suspension.

Some Pontiacs are equipped with engines produced by other GM divisions, subsidiaries, or affiliated companies worldwide. See your Pontiac dealer for details.

GM

MARK OF EXCELLENCE

PONTIAC WE BUILD EXCITEMENT

The world's densest compute is now the heart of the world's mo 32-bit computer.

Our 32-bit CPU.

From time to time, miracles of technology come along to make previously impossible tasks not only possible, but easy. That tiny 450,000-transistor integrated circuit is one of those technological miracles.

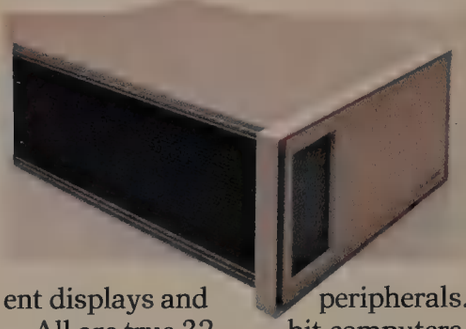
Hewlett-Packard didn't develop it just to break the record for most transistors on a chip, but to put on an engineer's or scientist's desk a computer so powerful that it can do the work of mainframes costing four times as much.

32-bit computers for 32-bit applications.

The new HP 9000 computer based on this and four other 'superchips' can handle formidable engineering and scientific problems. The scientist solving complex systems of equations, the mechanical engineer doing finite element analysis or three-dimensional modeling, the electrical engineer analyzing complex circuits or designing very large-scale integrated circuits — these are the kinds of technical people and problems the HP 9000 family is designed for.

It comes in three versions. The integrated workstation is complete with keyboard, color or monochromatic graphics display, fixed and flexible disc drives, and printer. For systems manufacturers, there's a rack-mountable box. And for a variety of single-user and multi-user applications, the minicabinet version works with many different displays and peripherals.

A rack-mountable version is available, too.



All are true 32-bit computers, with 32-bit CPUs, memories, and data paths. And the multi-CPU architecture lets you nearly double or triple your processing power at any time by adding one or two CPU boards. Without increasing the computer's size.

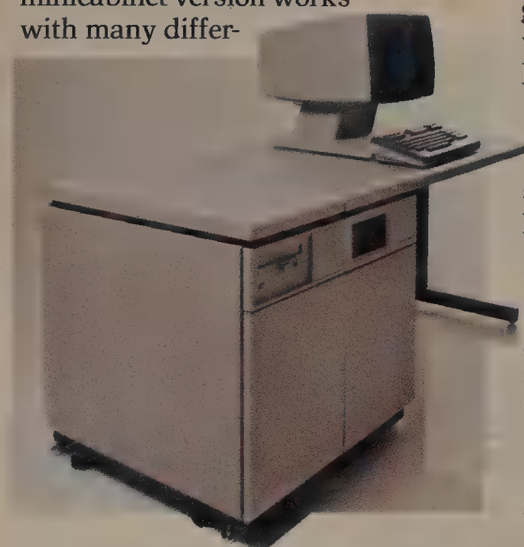
Two operating systems are better than one.

The integrated workstation is available with a choice of operating systems. One is HP's highly evolved, high-performance Enhanced BASIC, augmented with 3-D graphics and a software innovation called a run-time compiler. This substantially increases program execution speed, while retaining an interactive development environment.

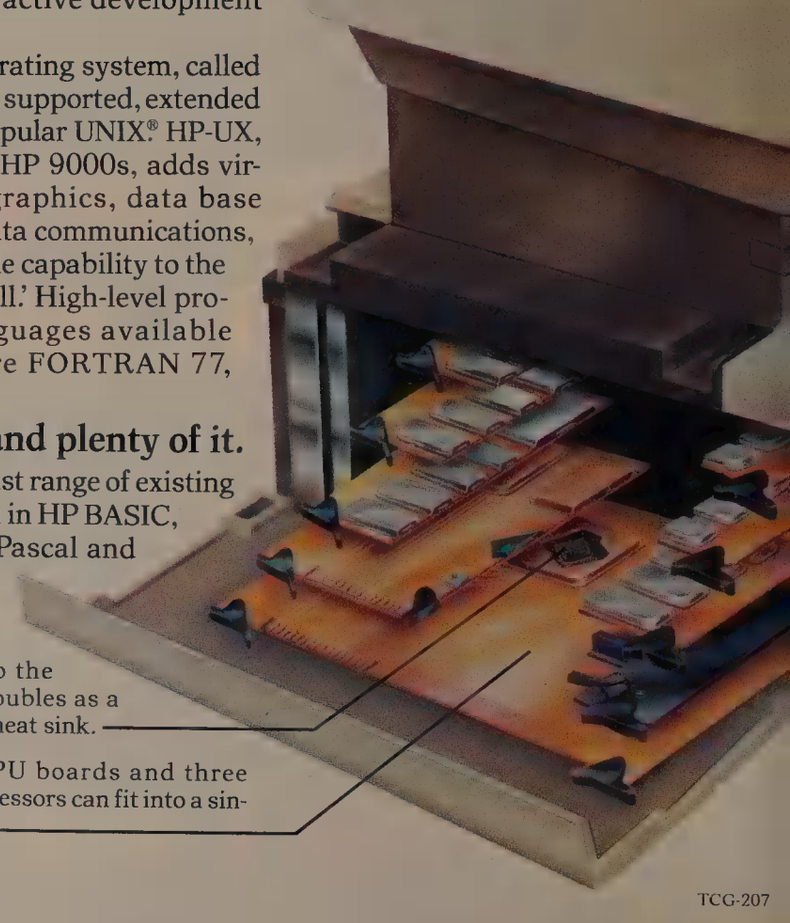
The other operating system, called HP-UX, is a fully supported, extended version of the popular UNIX® HP-UX, available on all HP 9000s, adds virtual memory, graphics, data base management, data communications, and enhanced file capability to the basic UNIX 'shell.' High-level programming languages available with HP-UX are FORTRAN 77, Pascal and C.

Software, and plenty of it.

Much of the vast range of existing software written in HP BASIC, FORTRAN 77, Pascal and C is transportable to the HP 9000. HP will also be offering proprietary software packages emphasizing computer-aided design and engineering. These will tie the HP 9000 into HP's Manufacturer's Productivity Network (MPN). Third-party software suppliers will be providing many of the most widely used CAE packages for 32-bit computer systems. And both HP 9000 operating



As a minicabinet, it can handle multiple users.



The 32-bit CPU chip is bonded to the finstrate which doubles as a signal carrier and heat sink.

Up to three CPU boards and three Input/Output Processors can fit into a single HP 9000.

TCG-207

hip
ffordable

systems offer extensive program development tools.

You also get a choice of communication tools. The HP 9000 is currently compatible with Ethernet,™ and with HP's Shared Resource Manager (SRM) which lets clusters of HP 9000 and 16-bit desktop computers share data and use common peripherals.

Links to central computers are also available. And in late 1983, HP will offer local area networks based on the IEEE-802 standard.

New technology from the silicon up.

The five superchips that make the HP 9000 possible are the 32-bit CPU, which can execute a million instructions per second; an eight-channel Input/Output processor (IOP); a random-access memory chip capable of storing 128K bits of data; a memory controller that 'heals' up to 32 bad memory locations; and an 18-megahertz clock.

Hewlett-Packard's advanced NMOS-III process makes it possible to put 450,000 transistors on a chip only 0.4 centimeters square. This tremendous density of electronic components could have required an expensive and elaborate cooling system.

Instead, HP engineers developed a new mounting structure called a finstrate, a copper-cored circuit board, which acts as both cooling fin and substrate. The finstrates containing the CPU, IOP, memory, and clock chips are housed in a lunchpail-sized module.

One user, one mainframe.

Clearly the trend in engineering and scientific computation is away from large machines shared by multiple users and towards networks of powerful personal workstations, sharing peripherals and data bases. The reason is compelling. An engineer or scientist in personal control of an HP 9000 can solve so many more problems more easily that the increased productivity alone makes the cost of individual computers easy to justify.

For complete information about this powerful breakthrough in 32-bit computing, contact the local HP sales office listed in your telephone directory. Ask a Technical Computer Specialist for a demonstration. Or write to Pete Hamilton, Dept. 56151, Hewlett-Packard, 3404 East Harmony Road, Fort Collins, CO 80525.

HEWLETT PACKARD

UNIX is a registered trademark of Bell Laboratories. Ethernet is a trademark of Xerox Corporation.

ic field are tightly frozen together, it is impossible to move plasma at right angles to the field without dragging magnetic field lines along. Thus a disturbance in the plasma can create a kink, or bulge, in the field lines. The kink can propagate along the magnetic field much as a kink in a stretched spring moves along the spring. Such a traveling disturbance is a magnetohydrodynamic wave, which carries energy as it displaces both the plasma and the magnetic field from their low-energy equilibrium states. As noted above, magnetohydrodynamic waves may be responsible for transporting energy from the convective layer under the solar photosphere, out through the photosphere and chromosphere and into the corona. Steepening into magnetohydrodynamic shock waves, the waves may dissipate their energy in the corona, generating the two-million-degree coronal temperature.

Why is there interest in coronal dynamics? Does coronal activity have any direct influence on the earth? If one understood the workings of the corona, one would have much insight into the complex behavior of hot, ionized gases in magnetic fields. One might then be able to understand the behavior of more bizarre and distant astrophysical objects, including pulsars, quasars and active galaxies. A better understanding of the corona might also contribute to the harnessing of nuclear fusion as an energy source, since most promising approaches to fusion involve the confinement of hot plasma by magnetic fields.

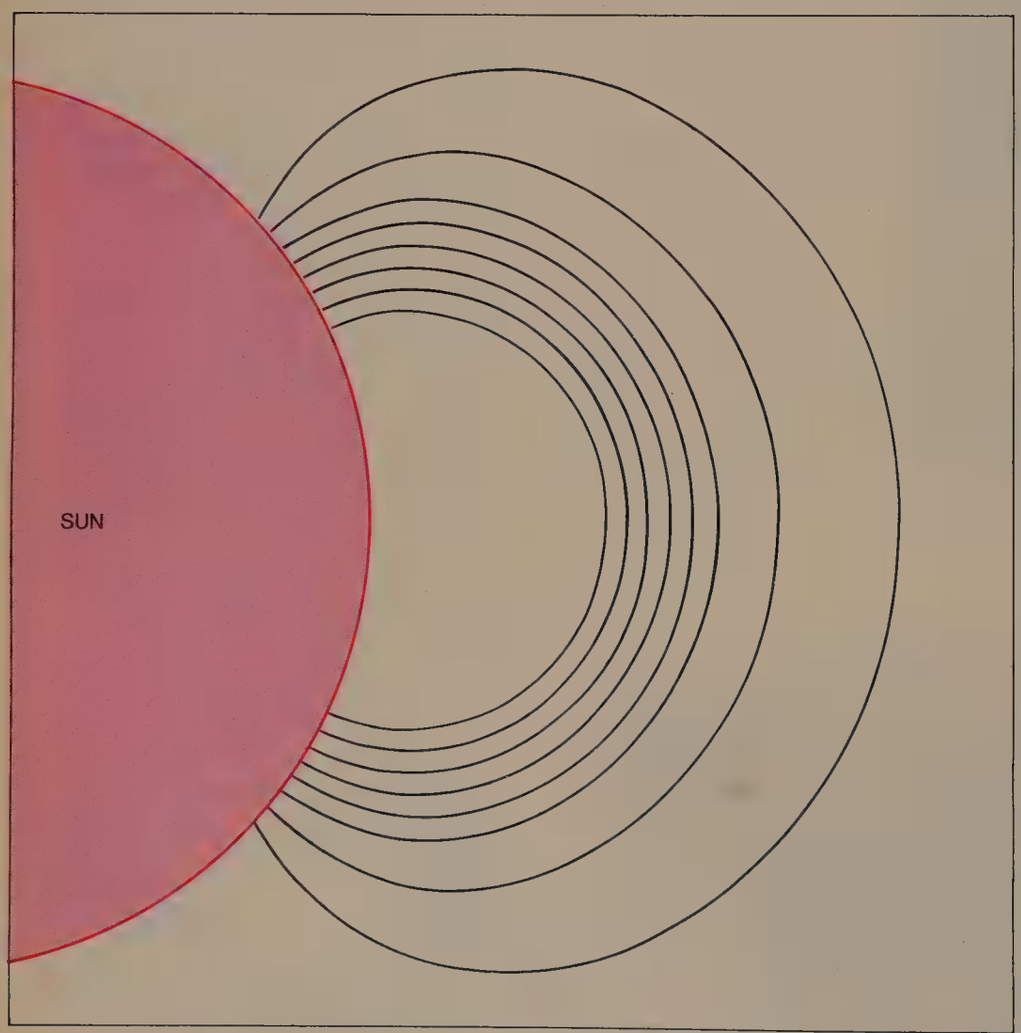
Of greater significance, however, is the fact that the earth is immersed in the corona. It has been recognized since the late 1950's that the high temperature of the corona is responsible for the solar wind: the outpouring of solar gas that rushes past the earth at some 400 kilometers per second. The solar wind is the extension of the corona into interplanetary space.

Since the coronal plasma is tightly gripped by the sun's magnetic field, parts of the corona in which magnetic loops are embedded cannot contribute significantly to the solar wind. With magnetic loops bound to the solar surface at both ends there is no way that coronal gas in the loops can escape into interplanetary space. In those regions where the lines of force in the sun's magnetic field stretch indefinitely outward, however, the solar wind can escape along the field lines. It is now thought that much of the solar wind, particularly its high-speed streams, originates in coronal holes. Coronal transient events may provide another avenue for material to escape from the sun altogether. In many transients an initially closed loop becomes unstable, rising and expanding to the point where its field lines open up, ejecting coronal material into space. Recent observations have also revealed the existence of "coronal bullets": knots of cool, dense material that accelerate rapidly through the corona and may contribute to the outflow from the sun.

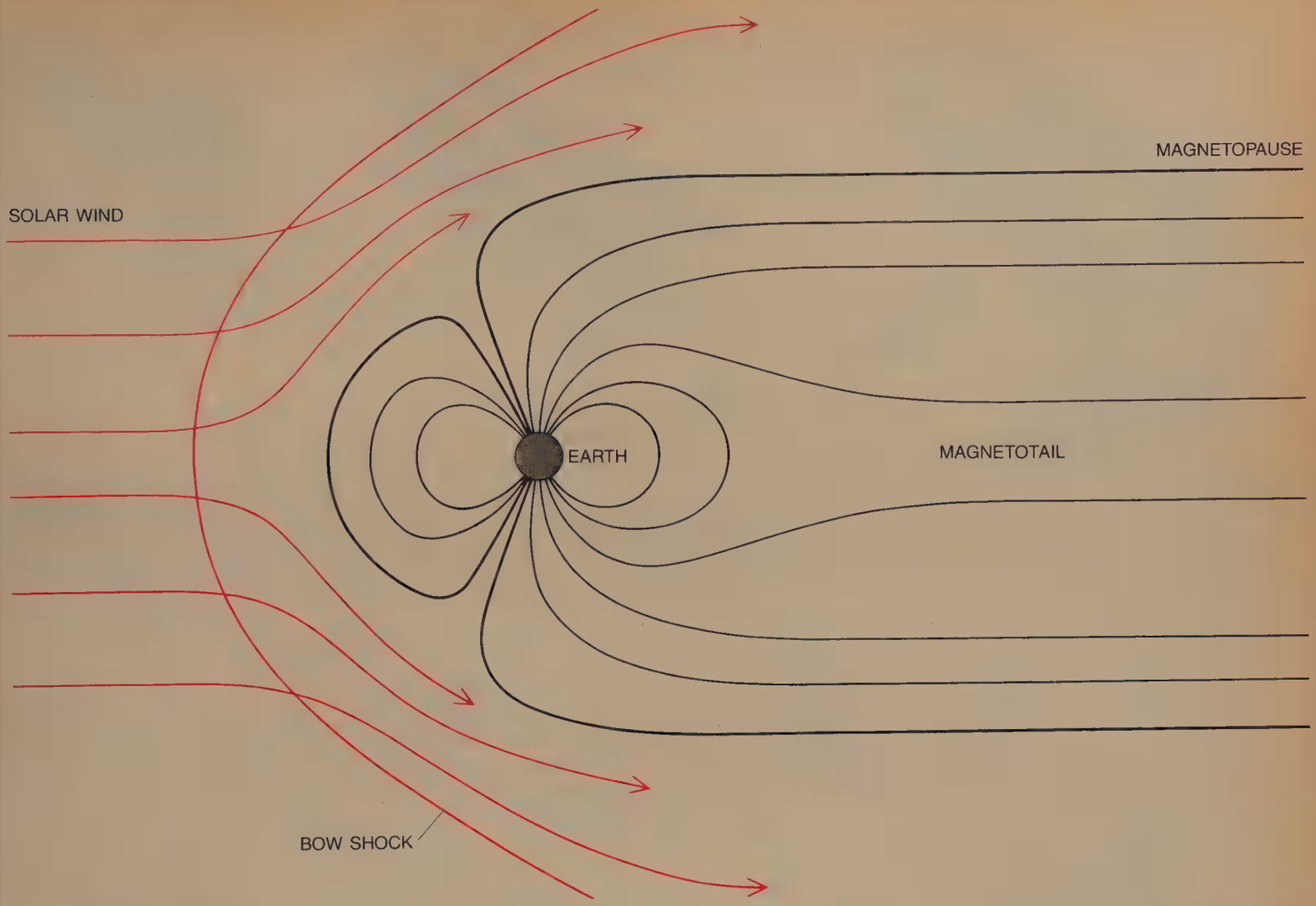
The earth's magnetic field shields the earth from the direct impact of the solar wind, but the wind distorts the field. In the direction toward the sun the earth's field is compressed; in the direction away from the sun the field is drawn out into a long tail. The region in which the earth's magnetic field is confined by the solar wind is termed the magnetosphere. Like the corona itself, the magnetospheres of the earth and other planets are laboratories for studying the interactions of plasmas and magnetic fields.

A magnetosphere is an obstacle in the path of the solar wind. Because the wind flows faster than the speed of sound or of magnetohydrodynamic waves, no signals can travel upstream from the magnetosphere to cause the solar wind to adjust to the obstacle. The situation is the same as when an airplane travels faster than sound or a boat moves faster than water waves. In all such cases the fluid adjusts suddenly and dramatically to the obstacle, rather than gradually as it would if its speed were less than that of sound or some other kind of wave. This sudden adjustment is a shock wave, where fluid properties such as pressure, density and velocity change abruptly. The shock wave created by a supersonic airplane is a sonic boom; that of a boat is a bow wave. A huge bow shock stands in space ahead of the earth's magnetosphere, about 60,000 kilometers from the earth. Spacecraft crossing the bow shock can gather valuable information about the structure of shock waves in plasmas.

If the solar wind were unvarying, the magnetosphere and the bow shock would remain at fixed locations determined by a balance between the total pressure of the incoming solar wind and the pressure of the earth's magnetic field. The wind, however, is an extension



DEVELOPMENT OF CORONAL-LOOP TRANSIENT is plotted according to a mathematical model recently devised by the author. The curves represent successive positions of an expanding loop of coronal material embedded in a magnetic field. The top of the loop ascends as the feet of the loop spread apart, perhaps in response to a violent event such as a solar flare or as a result of random motions in the gas at the surface of the sun. The model predicts that the rate of ascent of the loop is progressively faster than the lateral displacement of the feet of the loop. This result indicates an unstable situation where relatively small disturbances at the surface of the sun can initiate rapid changes in the structure of the corona. The model includes the effects of gas pressure, magnetic field and gravity. For mathematical tractability the magnetic interaction is assumed to be confined to thin layers coincident with the plotted curves.



PRESSURE OF THE SOLAR WIND confines the earth's magnetic field to the region called the magnetosphere, which is compressed in the direction facing the sun and drawn out into a long magnetotail in the direction away from the sun. The magnetopause, the boundary between the solar wind and the magnetosphere, acts as a barrier that prevents most of the particles in the solar wind from reaching the earth. Near the earth's poles, however, particles can move along magnetic field lines into the upper atmosphere, giving rise to auroras. Ahead of the magnetosphere is a bow shock that is analogous to the shock wave that generates the sonic boom of a supersonic aircraft.

of the corona, and the corona is itself highly variable in both space and time. As the sun rotates, the spatial irregularity of the corona—here a coronal hole, there a magnetic loop—translates into temporal variations in the solar wind at the distance of the earth. Dynamic events on the sun, such as photospheric flares or coronal transients, also give rise to variations in the solar wind.

In response to changing solar-wind pressure the earth's bow shock and magnetosphere move about, compressing the earth's magnetic field and altering its strength. Changes in the magnetic field induce electric currents in the ionosphere, the conducting layer of the upper atmosphere. This in turn can lead to sudden disturbances of long-distance radio communication. Similar currents can be induced in telephone and power lines, both buried and aboveground, and in the earth itself. Strong disturbances on the sun can therefore result in major disruptions of compass readings, communications and even electric-power transmission. At such times high-energy particles spiral down magnetic field lines at high latitude and give rise to auroras.

Streams of high-velocity solar wind emerge from coronal holes. As the high-speed wind pushes into the slower wind ahead of it, the transition zone between the fast and the slow wind is squeezed ever thinner and the transition becomes more abrupt. By the time the solar wind reaches the orbit of Jupiter, five times farther from the sun than the orbit of the earth, the transitions have steepened into hammerlike shock waves that pound Jupiter's magnetic field. Jupiter's magnetosphere fills a volume so large that even at its great distance from the earth it would appear larger than the sun if we could see it. In response to the pounding of the solar wind it can vary in size by a factor of three in a matter of days. Pioneer and Voyager spacecraft traveling near Jupiter crossed the planet's bow shock and magnetopause repeatedly, indicating that the two structures move rapidly back and forth, now swelling, now contracting.

If the corona is still much in evidence at Jupiter, where does it end? What is the full range of the sun's influence? At some point the decreasing pressure of the solar wind must become weaker than the minuscule pressures of the interstellar gas and the magnetic field of the galaxy. This transition region marks the heliopause: the end of the sun's physical extent. Theoretical calculations a decade ago suggested that the heliopause might be between the orbits of Jupiter and Saturn. With active spacecraft nearing the orbit of Neptune, no sign of the heliopause has been found. It is now believed the solar wind extends well beyond the planetary limits of the solar system.

Beyond the heliopause is a vast bow shock caused by the motion of the entire solar system through the interstellar medium. Driven by variability in the solar corona, the heliosphere, the region within the heliopause, may expand and contract like a balloon. Everything from the thermonuclear furnace at the sun's core out to the heliopause—through the convective zone, the photosphere, the chromosphere, the corona and the solar wind—is a component of the sun. Human beings occupy a privileged position within this giant celestial object, far enough from its core for comfort but close enough to be able to begin to understand the complex dynamics of the extended corona.

Dart-Poison Frogs

In Colombia, Indian hunters poison blowgun darts with highly toxic alkaloids secreted by small frogs. Both the alkaloids and the evolutionary biology of the frogs raise interesting questions

by Charles W. Myers and John W. Daly

Poisonous organisms are so numerous that chemists and pharmacologists are still many years away from analyzing all the noxious molecules found in various microorganisms, plants and animals. Basic research in this area is of interest to the broader fields of ecology and evolutionary biology; it may also lead to new drugs or other useful substances. Some particularly intriguing subjects for investigation are provided by the richness of life in tropical regions. An example in the New World Tropics is the plant genus *Strychnos,* a source of the poison strychnine and of the curare alkaloids. Curare is used by Indian hunters for poisoning arrows and blowgun darts and by anesthesiologists as a muscle relaxant. Another example is provided by a family of frogs, certain species of which secrete one of the strongest animal poisons known. This substance too is used as a dart poison and has found a place in biomedical research. Other species in the same family secrete many additional toxins, and the frogs themselves are also remarkable in several aspects of their biology.

The poisonous frogs and their nontoxic relatives belong to the family Dendrobatidae, which is confined to South America and southern Central America. They occupy a wide range of habitats. Some species are found along streams; others live away from water, on or near the ground in lowland or montane rain forest. A few forest species even spend most of their lives up in the trees. At the other extreme are a few species that live in open dry country, where they find sufficient humidity on shaded ground under low vegetation. In spite of this ecological diversity the dendrobatid frogs share a life style that sets them apart from virtually all other frogs. They are active only in the daytime, and they lay their eggs in moist places on land. They tend their eggs until hatching, when a "nurse" frog carries the tadpoles, literally glued to its back, to a suitable aquatic environment.

The dendrobatid family includes more than 100 species, currently divided among four genera. They are small frogs, varying in head-and-body length from little more than a centimeter to about five centimeters. Whereas among most frogs the male is smaller than the female, in many populations of dendrobatids the male is about as large as the female. Large males are associated with a high degree of territoriality and aggressiveness. The males of such species invest much time and energy in advertising their territory and attracting mates by prolonged calling, and in actively challenging and grappling with other males. Many territorial males are noisy little creatures; their insectlike chirps, peeps or trills contribute noticeably to the daytime sounds of some tropical forests. The sight of two miniature frogs wrestling on a leaf, chirping vigorously at each other, can be amusing to a human observer, but sexual fitness and successful breeding are evolutionarily serious matters. In some species of dendrobatids the female is also aggressive and defends a territory.

The relatively large size of male dendrobatids is also associated with unusual kinds of mating behavior. Among most frogs courtship concludes with the male mounting the larger female and clasping his mate either behind her forelegs or in front of her hind legs. Male dendrobatids clasp the female not around her trunk but around her head, with the backs of the fingers pressed up under the female's chin. This forward position better ensures the fertilization of the eggs being extruded by the female; if a male of similar size were farther to the rear, its sperm could fall to the ground. In some species of dendrobatids clasping is absent from the courtship ritual, and the eggs are fertilized in the course of rather complicated maneuvering by both the female and the male.

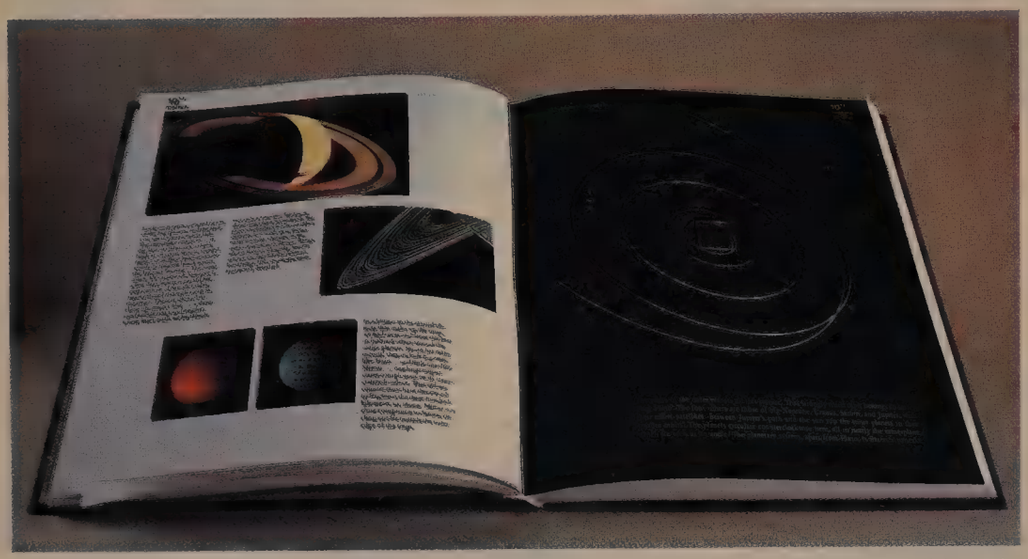
Depending on the species, dendrobatid eggs are laid in leaf litter, in crevices under rocks, on shaded leaves aboveground or in such tropical forest growths as bromeliads and arums. The size of the egg clutch is small compared with the hundreds or even thousands of eggs laid by many water-breeding frogs; it can be as few as one or two eggs among the smallest dendrobatids and is rarely more than 30 or 40. In many dendrobatid populations a small clutch size is partially offset by the fact that breeding is continuous throughout the year. Some individual frogs can breed every month. Statistics on this point are provided by observations of a pair of frogs of the species *Dendrobates tricolor* kept at the American Museum of Natural History for more than two years. These frogs mated from two to four times a month. Clutches were produced every 10th day on the average and consisted of from 12 to 30 eggs. Even with this prodigious effort the female laid only about 600 eggs each year.

Once they are deposited the eggs may be "guarded" by either parent or may be left alone except for periodic brief visits. On such a visit the frog may moisten the eggs, apparently with water from its bladder. The nurse frog occasionally wriggles its hindquarters into the egg mass, an action that ultimately aids the hatching process and helps the tadpoles to squirm up onto their parent's back. There they adhere to a patch of mucus secreted by glands in the frog's skin. Adhesion is facilitated by the flat or slightly concave belly that is characteristic of dendrobatid tadpoles. In some species the attachment is no more than superficial; in others the mucus is gluelike and the tadpoles are firmly fixed to the adult's back. Again depending on the species, the tadpoles may remain attached to the nurse frog for an interval ranging from a few hours to more than a week. In this period the tadpoles may grow to some extent by absorbing yolk reserves.

The nurse frog eventually travels to a suitable body of water, where a period of soaking loosens the mucous bond and the tadpoles swim free. Many dendrobatids take their tadpoles to small streams. Other species release them in the small

POISONOUS FROGS of the genus *Phyllobates* are shown 1.7 times life size in these paintings by David M. Dennis. The first two frogs at the top, from the left, are *P. lugubris* and *P. vittatus,* two Central American species that secrete relatively small amounts of toxins from glands in their skin. The striped frog at the upper right is *P. aurotaenia,* one of three Colombian species used for poisoning blowgun darts. In the middle row are two specimens of *P. bicolor,* the second of the three Colombian species. They show extremes in size and color variation; the legs, of a hue different from the body, vary from pale green to black. At the bottom is the third and most poisonous Colombian species, *P. terribilis.* The immature *P. terribilis* (*left*) is striped like its relatives at the top, but its stripes are being obliterated by the spread of bright pigment. The adult *P. terribilis* (*right*) may be yellow, orange or pale green, but normally it is uniformly colored above and below. *P. terribilis* is a bolder, less secretive frog than the other species and is at least 20 times as toxic. It is potentially dangerous even to handle.

SCIENTIFIC AMERICAN LIBRARY



MAN'S CORNER OF THE UNIVERSE lies open here at the pages in *Powers of Ten* that display the universe at the 10^{13} meters dimension (10 billion kilometers), about the diameter of the solar system.

POWERS OF TEN

A wondrous journey across the universe from its largest dimension to its smallest

Here is an atlas of the universe and a handbook to its exploration by the disciplines of science. In vivid color illustration—from the informed imagination and from the instruments of science—*Powers of Ten* shows how the universe looks from the largest to the smallest dimensions now in the grasp of human understanding.

This spectacular journey, with its sweeping changes in perspective, proceeds in ordered steps, each a multiple of ten of the last step or of the next.

There are 42 of these equal logarithmic steps from the outermost regions to the innermost recesses of the known universe. More than half were taken in the lifetime of scientists now alive.

The visual magic of this first-time-in-history journey is the bold invention of Charles and Ray Eames, who have here turned their mastery of multi-media technology to book design. Your tour guides are Philip and Phylis Morrison. They supply instructive and poetic pages of commentary, answering questions and showing how the sciences get those answers at each of the 42 steps.

The SCIENTIFIC AMERICAN LIBRARY publishes *Powers of Ten* as its premier volume because it sets out the agenda of the books that are to come. Each succeeding book further illuminates the ideas and images displayed, step by step, on the pages of *Powers of Ten*.

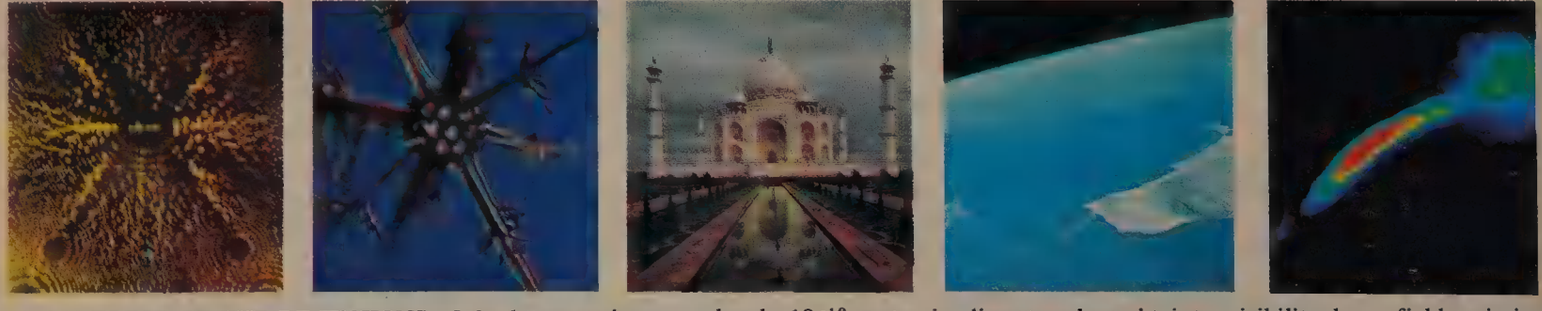
The SCIENTIFIC AMERICAN Library opens an entirely new approach to understanding and sharing in the work of science. Each specially commissioned volume

. . . comprehends a large topic that engages more than one discipline and, typically, recurs in many articles in SCIENTIFIC AMERICAN;

. . . is written by a scientist who has made his own pivotal contribution to the work on the topic;

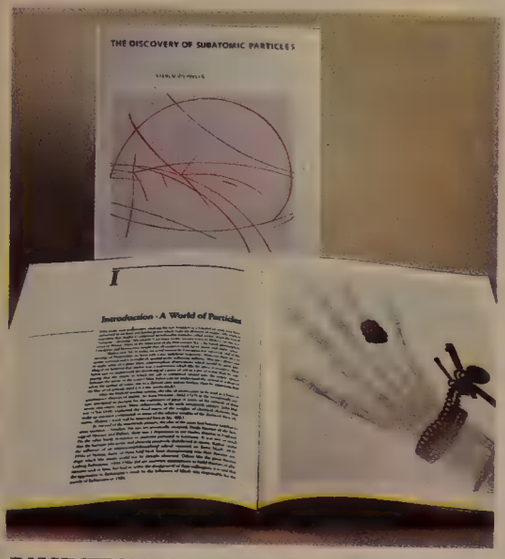
. . . rewards its reader not with "information" but, thanks to the sympathetic collaboration of the author and editor in the preparation of text and illustration, with command of the concepts central to understanding of the work and appreciation of its meaning.

Readers of this magazine are invited to accept membership in the SCIENTIFIC AMERICAN Library. Members are entitled to receive for inspection, without any charge, each volume as it is published. Membership, cancellable at any time, will gather in time a reference library of the member's own choosing, with a reach and accessibility unprecedented in the literature of science.



A SAMPLING OF THE DIMENSIONS of the known universe explored on the journey through the Powers of Ten is shown here. At center, the Taj Mahal adorns the 10 meter world in reach of our unaided senses. Dimensions grow smaller to the left: first, crystalline skeletons of single-celled diatoms at 10^{-4} meter (.1 millimeter) and at far left, tungsten atoms, 10^{-10} meter in diameter, brought into visibility by a field-emission microscope. At right, looking across the Indian Ocean toward the South Pole through the window of Gemini II, the curvature of the Earth, 10^{7} meters in diameter. At far right, a quasar, 10^{25} meters away (one billion light-years), caught by a radio telescope and imaged by computer graphics.

SUBATOMIC PARTICLES
by Steven Weinberg



DISSECTION OF ATOM began with accidental discovery of X-rays by Roentgen in 1895.

As everyone knows, the once irreducible atom has been shown to be composed of still more fundamental particles: the electron, proton and neutron. But how do we know this? In *The Discovery of Subatomic Particles* the reader comes to know the answer just as Steven Weinberg and other physicists did, thereby sharing in the intellectual enterprise that has reshaped physics in the twentieth century.

We do not accept scientific truth on faith or authority. We know truths of this kind are grounded on experiment and observation—operations we might have performed ourselves, but in any case can perform inside our heads. With Weinberg the reader of this book reenacts the historic succession of experiments that disassembled the atom. From each experiment to the next, the reader will gain a confident understanding of the laws of physics that govern the events encountered.

HUMAN DIVERSITY
by Richard C. Lewontin

The relevance of science to human values finds compelling demonstrations in the proper understanding of the genetics of our species. In *Human Diversity* Richard Lewontin shows that each human being differs from all others because of the interaction of genetic differences, environmental differences, and chance events that occur during development. There is no unique "normal" genetic constitution.

The genetic markers that supposedly divide the species into races are only a trivial fraction of the total spectrum of diversity in which each person finds singularity. The genetic endowment held in common across racial, national and cultural boundaries is a consequence of the biological history of our species. The segregation of people into social castes can never again invoke genetics as its rationalization.

FOSSILS
by George Gaylord Simpson

One of the architects of the modern evolutionary synthesis here establishes the nature—and the true splendor—of the solid evidence upon which much of the theory and the facts of evolution rest. Simpson declares: "The primary record of the history of life is written in the successive strata of rocks as in the pages of a book. Fossils may be called the writing on those pages. Traces of organisms living at successive geological times, they represent once living things and should be seen as such. They must be put in their sequence in time. They were influenced by and bear witness to geographic and geological changes on the earth. They are basic materials for the study of organic evolution. Their study combines historical geology and historical biology into one great synthesis."

THE SOLAR SYSTEM
by Roman Smoluchowski

The fine-grained picture of the Earth and its immediate cosmic neighborhood now in the possession of mankind is here comprehensively assembled. The author has brought together the latest graphic images from the satellites and from earthbound telescopes as well. (We know the terrain of cloud-wrapped Venus from radar echoes returned to the 600-foot radio telescope at Arecibo in Puerto Rico.) He shows us our solar system in time as well as in space. In the wealth of images of the Sun, of the planets, of their satellites, of the swarming asteroids and of the cometary messengers from the primordial ice clouds on the outermost circumference of the system, we comprehend the origin, the history and the fated future of our Sun and its planets.

MUSICAL SOUND
by John R. Pierce

In this volume, the author will share with you his joy in the discovery that understanding from physics and mathematics enlarges and enhances one's appreciation of music. Thereupon you will find yourself in the company of a long succession of natural philosophers who have pursued just that strategy.

John R. Pierce is the latest of these natural philosophers. To his work in music he brings his study of sound, its transmission and reproduction, at Bell Telephone Laboratories plus the insights of such contemporary composers as Pierre Boulez, John Cage and Edgard Varèse.

In the live recordings that come with this volume Pierce demonstrates something of what the psychology of acoustics has learned about the perception, the illusion and the effect of sound. He imbues the reader with his confidence that rational enquiry into this most intensely subjective of all aesthetic experiences will open up new realms of sound to that experience.

SOUND SHEETS, for 33-r.p.m. turntable, demonstrate acoustic illusions discussed in book.

Charter Membership Reservation Form

To SCIENTIFIC AMERICAN LIBRARY
Holmes, Pennsylvania 19043

Please enroll me as a Charter Member of the SCIENTIFIC AMERICAN LIBRARY. Send me the Premier Volume — *Powers of Ten* — by return mail. Thereafter, I am entitled to receive for inspection ■ new volume every two months, as it is published, for fifteen day's *free examination*.

I need send no payment now. For each volume I keep I will be billed $19.95 plus postage and handling. This price is guaranteed for at least one year, regardless of any price increases asked of new members.

I understand that I may cancel my Charter Membership at any time for any reason.

NAME _____
(Please print clearly)

ADDRESS _____

CITY _____ STATE _____ ZIP _____

Canadian residents: Please pay in U.S. dollars.

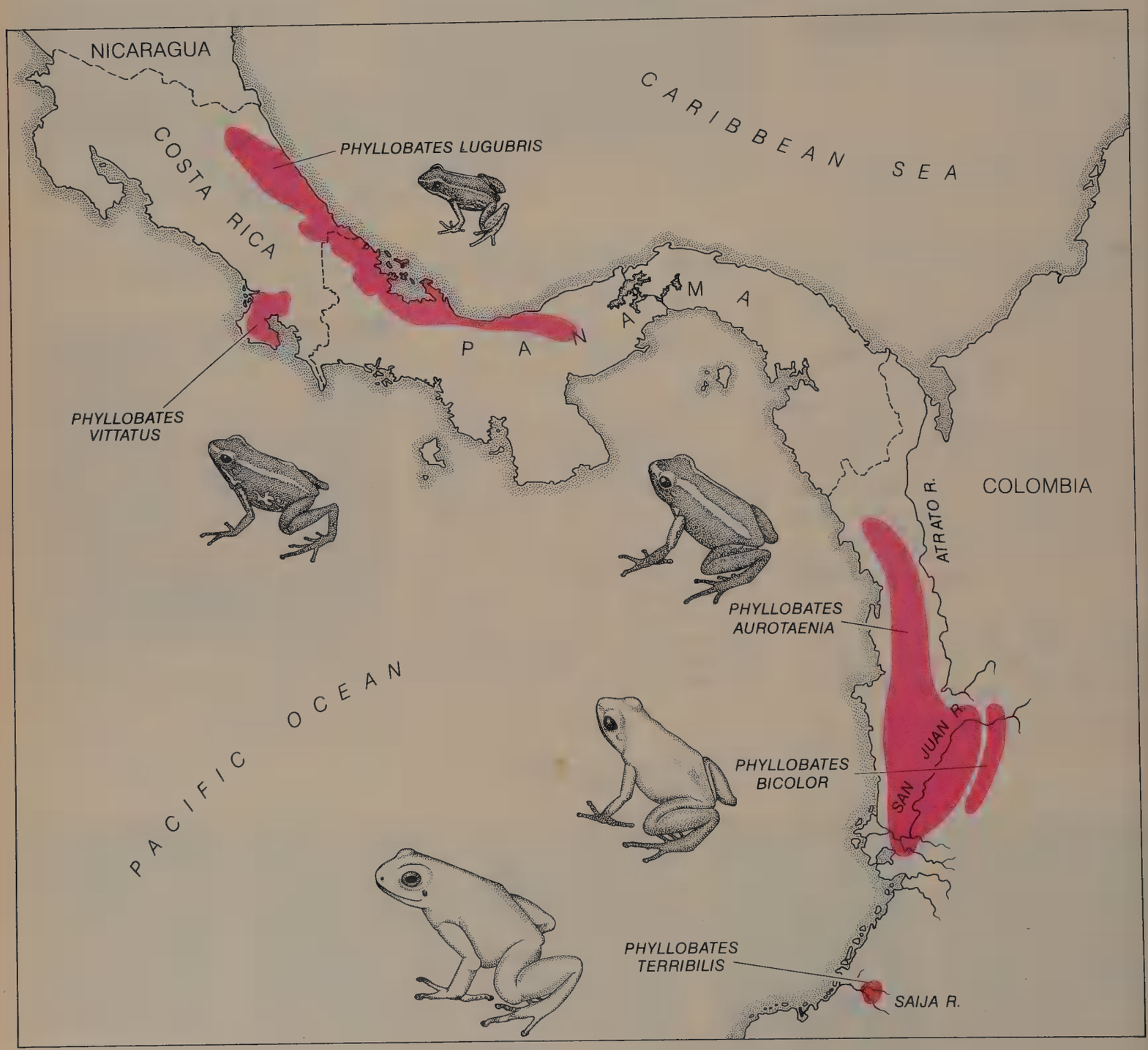
FD

pools of rainwater that collect in bromeliads or between the leaf and the stem of certain other tropical plants. The nurse frogs of one species, *Dendrobates auratus,* often release their tadpoles in water retained by hollows in tree trunks. The tadpoles of *D. auratus* tend to resort to cannibalism in this nutrient-poor environment, and only one of them may survive to maturity.

Rainwater in small bromeliads and in the axils of leaves also seems to hold few nutrients, a circumstance that helps to explain a surprising observation made recently by Peter Weygoldt at the University of Freiburg in West Germany. Working with laboratory specimens of *Dendrobates pumilio,* a species that releases its tadpoles in such places, Weygoldt found that the females regularly visit their free-living tadpoles to deposit unfertilized eggs in the water. The tadpoles then feed on the eggs!

Biologists sometimes attempt to determine the relative "cost" of parental care by male or female animals, because such care may impose limitations on reproduction or may expose the parent to predation. These are factors to be weighed against the benefit of ensuring success for one or a few sets of eggs or for sperm from an individual parent. Among dendrobatid frogs the reproductive cost may be higher in species (such as *D. pumilio*) where the nurse frogs are usually females than it is in species where the nurse frogs are usually males. Of the *D. tricolor* pair at the American Museum the male was able to care for several clutches of eggs at a time. Furthermore, it vocalized frequently and presumably would have mated with other females if they had been available. It fed even while it was carrying on its back a full load of tadpoles, which were released into the water within hours of hatching.

It has been supposed that the sex of the nurse frog (male in most species and females in some) is a species-specific trait. In a few species, however, both sexes have been found carrying tadpoles, a fact that is hard to explain. How can such a novel behavioral trait be so variable? Experiments with the *D. tricolor* pair suggest that the answer may involve the outcome of sexual competition for the nurse role. The female of the pair was both able and willing to



DISTRIBUTION OF THE FIVE SPECIES of the dendrobatid genus *Phyllobates* is patchy. The three species found in western Colombia are the only frogs of any kind known to be used by Indian hunters for poisoning blowgun darts. The two species found in Costa Rica and Panama secrete small amounts of the same toxins as their Colombian relatives, but they do not inhabit the areas where blowguns are used.

be the responsible parent but was kept from the role by the territorial aggressiveness of the male. The female stayed with her clutch of eggs for the first hour or so; thereafter she was apparently allowed no further visitation rights. If, however, the male was removed after breeding, the female cared for the eggs and eventually carried the tadpoles as well. If the male was not removed until a few days later, the female showed no reluctance to cannibalize either her eggs or her tadpoles as though they belonged to another female. Such cannibalism may be a mechanism by which population density is regulated in some species of dendrobatids.

Much remains to be learned not only about the natural history of dendrobatids but also about their evolutionary relations and classification. The approximately 130 species are currently placed in four genera: *Atopophrynus, Colostethus, Dendrobates* and *Phyllobates. Atopophrynus* has only one species. *Colostethus* has more than 70, mainly brown in color. With a few exceptions these species are nontoxic. Most of the approximately 50 species of *Dendrobates* and the five species of *Phyllobates* are characterized by bright "warning" coloration. The brilliance of these frogs advertises the presence of poisonous or at least distasteful skin secretions that are effective in deterring many potential predators. Toxic dendrobatids are not, however, completely immune from predation; for example, they are preyed on successfully by large spiders and certain snakes. The frogs' secretions are released from microscopic glands in the skin at times of stress, and a predator that has seized a frog in its mouth will experience sensations of burning, numbing or foul taste that tend to make it drop the prey.

The most toxic dendrobatids are three species of *Phyllobates* from west of the Andes in the Pacific drainage of Colombia. These frogs secrete a poison that is much stronger than curare. Just as curare is a widely used poison for arrows and blowgun darts east of the Andes, so these dendrobatid toxins serve as dart poisons in western Colombia. Possibly the earliest report of their use is one written by Captain Charles Stuart Cochrane, who explored in Colombia in 1823–24 while he was on leave from the British navy.

Cochrane, crossing the western Andes on foot, had his attention drawn to frogs "called *rana de veneno* [by the Spanish], about three inches long, yellow on the back, with very large black eyes." "Those who use [their] poison catch the frogs in the woods, and confine them in a hollow cane, where they regularly feed them until they want the poison, when they take one of the unfortunate reptiles and pass a pointed piece of wood down his throat, and out at one of his legs. This torture makes the poor frog perspire very much, especially on the back, which becomes covered with white froth: this is the most powerful poison that he yields, and in this they dip or roll the points of their arrows, which will preserve their destructive power for a year. Afterwards, below this white substance, appears a yellow oil, which is carefully scraped off, and retains its deadly influence for four or six months, according to the goodness (as they say) of the frog. By this means, from one frog sufficient poison is obtained for about fifty arrows."

CARE OF THE YOUNG by the male or female parent includes attention to the eggs. Here a male *Dendrobates silverstonei* watches over a clutch deposited on a leaf on the forest floor.

Cochrane's *Travels in Colombia* seems not to have been widely read. A few later observers independently verified this curious use of frogs, however, and unsuccessful attempts were made to analyze samples of the poison scraped from Indian darts. The Swedish anthropologist S. Henry Wassén carried out field investigations in Colombia in 1934 and in 1955. Wassén determined that frog poison was used by two related Chocó Indian groups along the San Juan River. The frogs he obtained from the Indians, erroneously identified by zoologists as *Dendrobates tinctorius,* were actually *Phyllobates aurotaenia* and *P. bicolor.* Wassén also got some frog-poison darts from an isolated group of Chocó far to the south, on the Saija River. The frogs at this locality were collected by our colleague Borys Malkin and us in the early 1970's. Our specimens represented an undescribed species, which proved to contain at least 20 times more poison than its relatives along the San Juan. Because of the frog's awesome toxicity, we gave it a name with obvious connotations: *Phyllobates terribilis.*

The northern Chocó Indians poison their darts with the secretions of *P. bicolor* and *P. aurotaenia* much as described by Cochrane a century and a half ago. The frogs are impaled on a stick and are sometimes held near a fire before the darts are poisoned. The southern Chocó Indians, however, poison their darts simply by wiping them across the

"NURSE" FROGS carry their tadpoles to a suitable body of water after hatching. A female of the species *Colostethus inguinalis* from Panama (*left*) is shown with a load of 27 tadpoles. A male of the small Peruvian species *Dendrobates reticulatus* (*right*) carries only one. The young may remain attached to the parent in this manner for a few hours or for more than a week.

DATSUN 200-SX

IT'S GOT PERFORMANCE DOWN TO A SCIENCE.

The sleek Datsun 200-SX is a scientific paradox. Although it burns gas "faster" than conventionally powered cars, the 200-SX gets even better gas-mileage ratings this year than last. And last year's ratings were already remarkably high for a car that won two dozen races in just nine months.

How? The 2.2-liter Nissan NAPS-Z engine. A microprocessor continuously adjusts the fuel/air mixture injected into the engine's crossflow hemihead combustion chambers and selects which of each cylinder's twin sparkplugs to fire, a unique concept in engine design.

Combustion in this "fast-burn" system is completed within 80° of crankshaft rotation. Ordinary engines don't complete combustion until 170°, wasting both power and fuel. The Nissan NAPS-Z saves both by burning gas "faster."

Most important, you can have this scientific advance without compromising looks or luxury features. This unique power plant delivers its high performance wrapped in a sculptured body that turns science into art. See and test drive the 200-SX at your Datsun dealer now.

38 EPA EST HWY (27) EPA EST MPG

Use EPA estimated mpg for comparison, with standard 5-speed. Actual mpg may differ depending on speed, trip length and weather. Highway mpg will probably be less.

BELT YOURSELF

DATSUN
WE ARE DRIVEN

THE NAME OF QUALITY
NISSAN

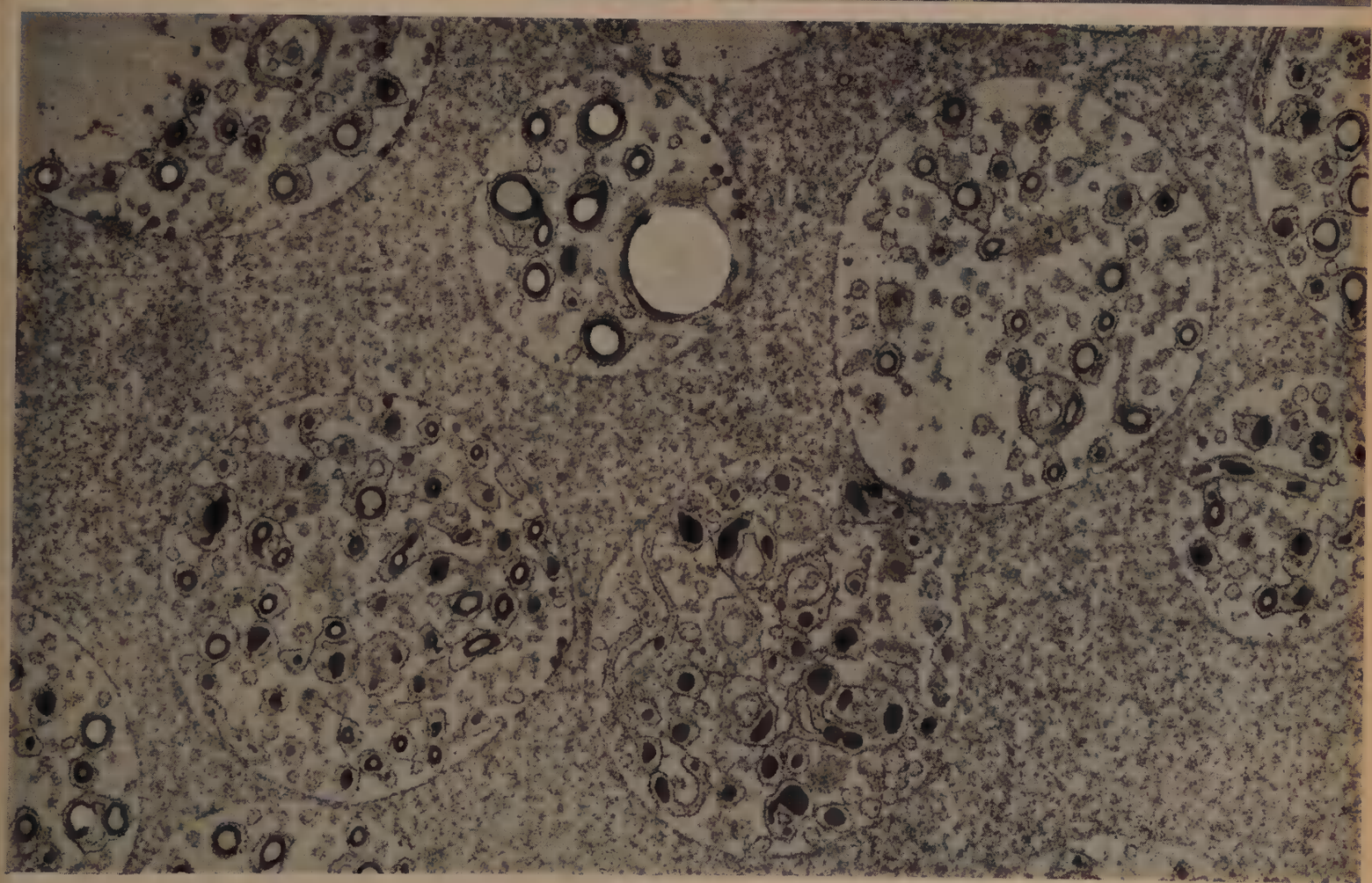
Product of NISSAN

©1983 NISSAN MOTOR CORPORATION IN USA

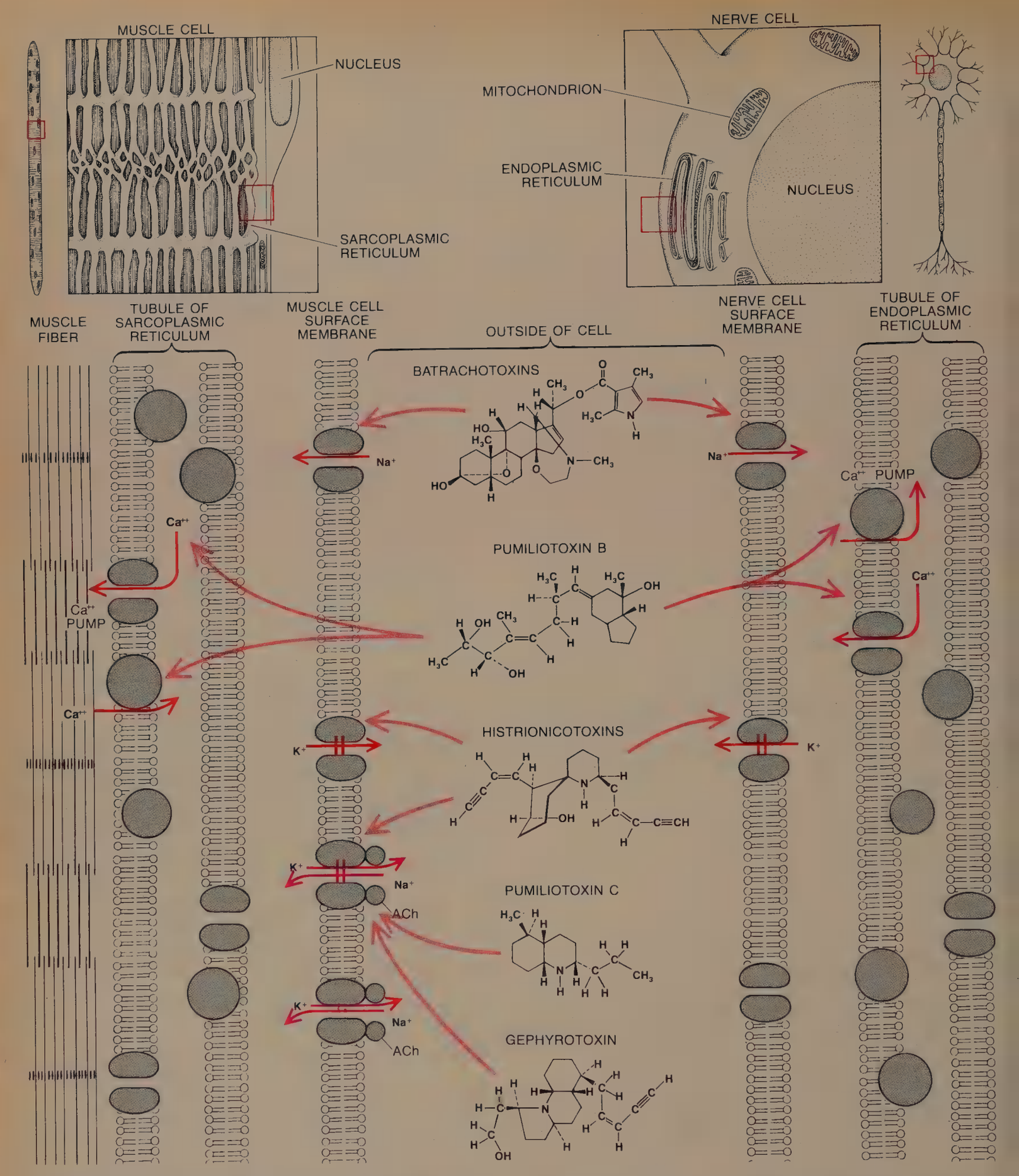
back of a living specimen of *P. terribilis.*

Today reliance on the blowgun is declining among the Chocó, mainly because of increased accessibility of firearms. Contrary to some literary accounts, dendrobatid frogs are not used for poisoning arrows, and their use for poisoning blowgun darts does not seem to have been widespread. This is understandable in view of the fact that the three most toxic dendrobatids are limited to relatively small areas of western Colombia. Two other species of *Phyllobates* are found in Central America, but they have much less toxin than their Colombian relatives and their range is outside the area where dart poisons are known to have been employed.

The skin organs that secrete the dendrobatid toxins are "granular" glands whose minute openings are scattered over the surface of the skin among the openings of the frog's mucous glands. Granular glands seem to be a primitive



"GRANULAR" GLANDS OF THE SKIN store and probably manufacture the toxic secretions of dendrobatid frogs. The photomicrograph at the top left shows a vertical section through the upper part of such ■ gland, magnified 900 times. The gland is filled with large, circular granules; the duct of the gland penetrates ■ layer of pigment cells and passes through the epidermis to open onto the surface of the skin. The electron micrograph at the top right shows ■ section through ■ gland, magnified 1,500 times. Vesicles in the granules may contain enzymes involved in the synthesis of the toxins. The electron micrograph at the bottom shows granules and the vesicles within them magnified 13,000 times. The micrograph at the top left shows ■ gland of *Dendrobates tricolor,* the other micrographs a gland of *D. auratus.*



MOLECULAR SITES OF ACTION for four kinds of alkaloid compounds from dendrobatid frogs are displayed schematically for muscle at the lower left and for nerve at the lower right. The morphological features of a muscle cell are shown at the upper left, those of a nerve cell at the upper right. In the middle are the structural formulas for the alkaloids. Batrachotoxin prevents the closing of sodium-ion channels in the surface membrane of both muscle and nerve cells. The influx of sodium ions electrically depolarizes the membrane and halts muscle and nerve function. Pumiliotoxin *B* appears to act in two ways. First, it facilitates the release of calcium ions from storage sites inside the muscle cell, thereby potentiating muscle contraction. Second, it inhibits the return of calcium ions to the storage sites, thereby prolonging the contraction. Pumiliotoxin *B* also affects the translocation of calcium in nerve cells. Histrionicotoxin blocks both the outward movement of potassium ions through potassium-ion channels of the surface membrane of muscle and nerve cells and the two-way exchange of sodium and potassium ions through complexes of ion channels and acetylcholine (ACh) receptors in the "end plate" between a nerve fiber and a muscle cell. The blockage of potassium-ion channels promotes the contraction of the muscle cells and prolongs the release of neurotransmitters by the nerve cells. The blockage of the ion-channel/acetylcholine-receptor complexes prevents the acetylcholine released from nerves from triggering muscle contraction. Both pumiliotoxin *C* and gephyrotoxin also block the movement of ions through the ion-channel/acetylcholine-receptor complexes, preventing acetylcholine from triggering muscle contraction.

characteristic common to all frogs. They were evidently a convenient evolutionary preadaptation in some groups of frogs for the synthesis, storage and release of various biologically active secretions.

The active principles of the dendrobatid skin secretions are alkaloids: nitrogenous ring compounds. Although such compounds are found more often in plants than they are in animals, dendrobatid frogs have proved to be a rich source of unique alkaloids. We have collected dendrobatids throughout their range in the tropical Americas for later analysis of their defensive secretions. This work has resulted in the discovery of a dozen new species of poison frogs and more than 200 new alkaloids, the latter representing at least five distinct classes of compounds. Chemical investigation of the frog alkaloids was initiated at the National Institute of Arthritis, Metabolism, and Digestive Diseases by Bernhard Witkop in the early 1960's. Over the past two decades major contributions to this research have been made by Isabella L. Karle, an X-ray crystallographer at the Naval Research Laboratory, and by Takashi Tokuyama, a chemist at Osaka University. The basis for the activity of each class of dendrobatid alkaloids is of considerable interest. The pharmacological investigation of these compounds has been pioneered by Edson X. Albuquerque of the University of Maryland School of Medicine.

The large majority of dendrobatid alkaloids include within their molecular structure a ring consisting of one nitrogen atom and five carbon atoms: a piperidine ring. Simple piperidine compounds are found among all subgroups of the genera *Dendrobates* and *Phyllobates*. From an evolutionary perspective this is evidence that the frogs in these genera form a separate lineage within the family Dendrobatidae. In the five species of the genus *Phyllobates,* however, the biosynthesis of piperidine alkaloids has been largely suppressed in favor of the synthesis of a new class of extraordinarily toxic alkaloids. They are the batrachotoxins (from the Greek *batrachos,* frog), which are the main basis of the poison on the Chocó blowgun darts. These complex alkaloids are structurally related to steroids but have many features hitherto unknown among natural compounds.

The batrachotoxins are among the most potent naturally occurring nonprotein toxins. They selectively increase the permeability of the outer membrane of nerve and muscle cells to sodium ions. This effect results in an irreversible electrical depolarization of the cells, causing in the heart arrhythmias, fibrillation and failure. The specific site at which batrachotoxin acts is associated with the channels that regulate the flow

Which of these languages would you like to speak?

Mark the one you want to speak in 2 or 3 months' time

Afrikaans	☐	French	☐	Japanese	☐
American English	☐	German	☐	Norwegian	☐
Arabic	☐	Greek (Modern)	☐	Polish	☐
Chinese	☐	Hebrew (Modern)	☐	Portuguese	☐
Czech	☐	Hindi	☐	Russian	☐
Danish	☐	Indonesian	☐	Spanish	☐
Dutch	☐	Irish	☐	Swedish	☐
Finnish	☐	Italian	☐	Welsh	☐

A Linguaphone Course makes you feel at home in almost any country. You'll never miss the meaning of conversations or be at a loss for words.

- It must work—over 4 million Linguaphone students in 88 countries speak a second language FLUENTLY.
- Proven learning success. You LISTEN to real conversations on cassettes...UNDERSTAND what you hear by following illustrated textbooks. HOLD CONVERSATIONS with the speakers.
- **You start speaking the very first lesson.** You learn at your own convenience.
- It's like having a private tutor.
- You get a complete, professional language program at little cost.
- You gain a good, working vocabulary.
- In just 2 to 3 months you can speak another language with complete confidence.
- You develop an authentic accent. Only native-born speakers are used.

Linguaphone The Language Masters

MONEY BACK GUARANTEE—28 DAY FREE TRIAL

World Language Courses, Inc. Dept. 227 313 Nolana Ave. McAllen, Texas 78501

FREE INFORMATION: Please mail me FREE information about learning the languages I have checked. FREE brochure.

Name (please print) _____

Street _____

City _____ **State** _____ **Zip** _____

SCIENTIFIC AMERICAN

is now available to the blind and physically handicapped on cassette tapes.

All inquiries should be made directly to RECORDED PERIODICALS, Division of Volunteer Services for the Blind, 919 Walnut Street, 8th Floor, Philadelphia, PA 19107.

ONLY the blind or handicapped should apply for this service. There is a nominal charge.

Free! Edmund Scientific Catalog

4,000 Unique Items

Astronomy, Microscopy, Biofeedback, Weather, Alternate Energy, Binoculars, Optics, Magnets, Magnifiers, Tools, Unique Lighting, Lab Equipment, and much more. Over 4,000 unique and fascinating products. Send for our FREE, colorful Edmund Scientific Catalog. Today!

Rush me your free catalog!

Name _____

Department _____

Institution _____

Address _____

City _____

State _____ Zip _____

Clip And Mail Coupon Today To:

Edmund Scientific, Dept. 8332 2110 Edscorp Bldg., Barrington, N.J. 08007

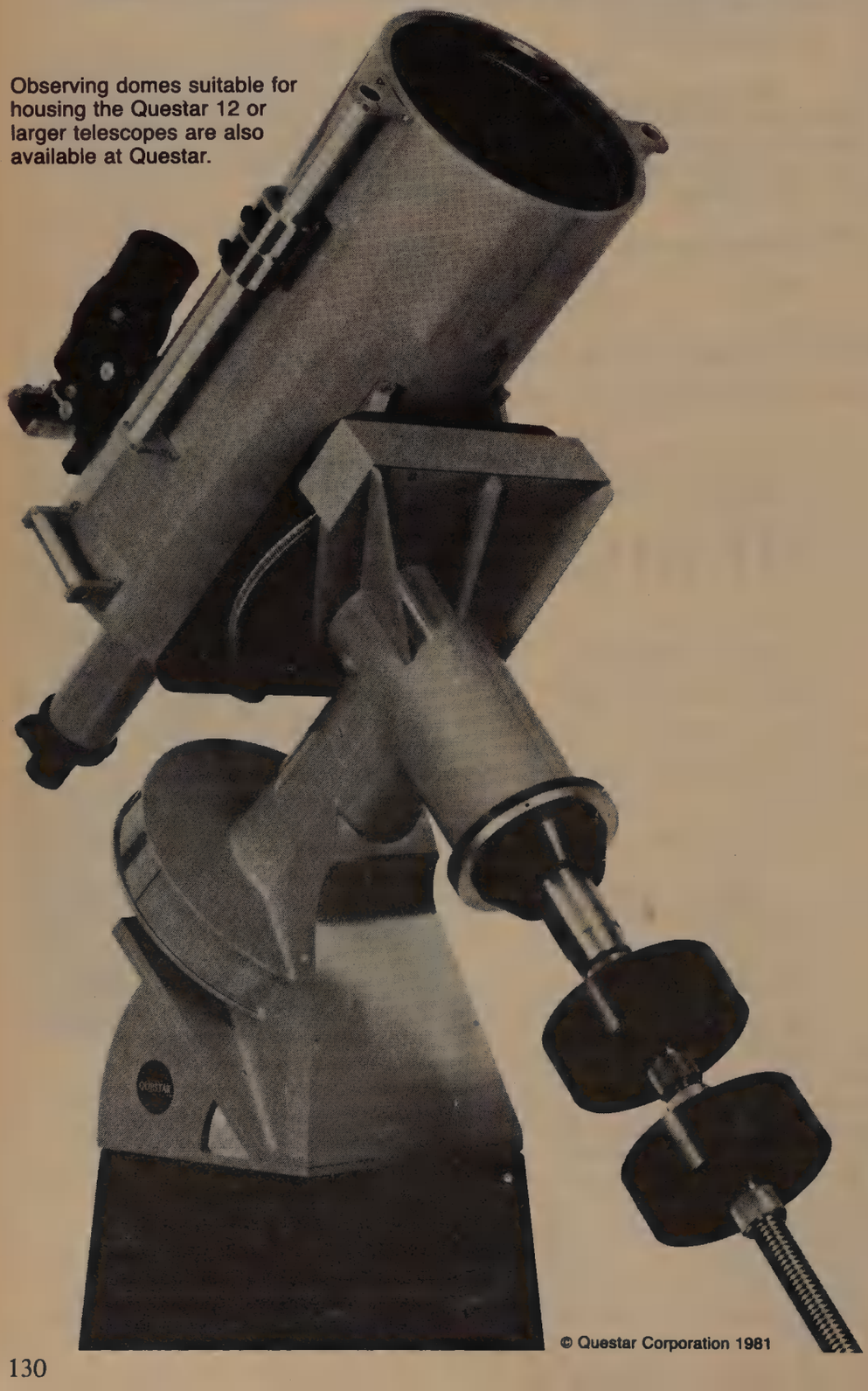
QUESTAR® 12 on the QUESTAR® MOUNT

The Questar 12, latest addition to the Questar family of fine optical instruments, now has its own Questar-designed mount. A German equatorial type, it is notable for its 360° continuous tracking in R.A. with precision tracking to better than 4 arc seconds. The Questar Mount is designed with over-sized components so that it can accommodate any Questar up to 18 inches. The standard mount shown is straightforward in design but can be modified so as to be compatible with more sophisticated tracking devices or other special equipment.

The Questar 12 is a superb instrument for the serious astronomer, for the university astronomy department or the engineer seeking sophisticated tracking and surveillance equipment for which Questar Corporation has a noted reputation.

Questar Corporation Box C, Dept. 20, New Hope, Pa. 18938 (215) 862-5277

Observing domes suitable for housing the Questar 12 or larger telescopes are also available at Questar.



© Questar Corporation 1981

of sodium ions through the cell membrane. Such sodium channels play a key part in the conduction of electrical impulses by nerve and muscle cells. The binding of batrachotoxin prevents the normal closing of the channels. As a result there is a massive influx of sodium ions and the cell is depolarized. Nerve cells can no longer transmit impulses and muscle cells remain in an activated, contracted state. The use of batrachotoxin in research has yielded important new knowledge of the function of sodium channels.

The other dendrobatid alkaloids are all much simpler in structure than batrachotoxin and are much less toxic, but they are no less interesting. Many of these piperidine alkaloids act on an ion channel in the "end plate" at the junction between a nerve fiber and a muscle cell. This channel is associated with a receptor for the neurotransmitter substance acetylcholine. The interaction of the neurotransmitter and its receptor opens the channel in the cell membrane through which both sodium and potassium ions can flow. If the resulting electrical impulse is large enough, it can trigger the opening of separate sodium channels in the membrane. With muscle cells this process causes the release of calcium ions inside the cell, which leads to contraction.

Histrionicotoxins, unusual spiropiperidine alkaloids first isolated from the frog *Dendrobates histrionicus,* interact with sites on the complex formed by the end-plate channel and the acetylcholine receptor, thereby blocking the passage of ions. Such blockage can prevent the transmission of signals from nerve to muscle, an end result similar to the one brought about by curare. Histrionicotoxins also block the passage of ions through separate potassium channels. When these channels are open, potassium ions flow out of both nerve and muscle cells, allowing them to return to their "resting" state after nerve transmission or muscle contraction. The blockage of such channels can lengthen the transmission of nerve messages and prolong muscle contraction.

Pumiliotoxin *B,* a unique indolizidine alkaloid first isolated from the frog *D. pumilio,* appears to affect the transport of calcium ions. In muscle cells an electrical impulse causes the release of calcium ions from internal storage sites. The ions then interact with intracellular proteins of muscle and cause contraction. Rather than blocking this process, pumiliotoxin *B* seems to facilitate it and also seems to delay the return of the calcium ions to their storage sites. As a result the force of contraction is augmented and contraction in both heart and skeletal muscle is prolonged. Such tonic actions on muscle may lead to clinical applications for this class of alkaloids.

Dendrobatid frogs with toxic or nox-

ious secretions are usually brightly colored. As is often the case with noxious animals, these species are mimicked by a few other species of frogs. An example is *Eleutherodactylus gaigeae,* which is found together with *Phyllobates lugubris* and *P. aurotaenia* and in the intervening territory, where an ancestral form of *Phyllobates* probably once lived. Unlike the dendrobatids it resembles, the nontoxic *E. gaigeae* is active at night. By day, however, it is sometimes found in the leaf litter where the *Phyllobates* species seek shelter when they are pursued. The different activity cycles may therefore be inconsequential as far as potential daytime predators are concerned.

The genus *Phyllobates* shows less variation than the genus *Dendrobates,* the species of which exhibit a remarkable variety of colors and patterns. Many *Dendrobates* species have primitive striped patterns, but others have evolved colorations that are uniformly bright or variously spotted, banded or mottled. Such variation may even occur among populations of the same species. It would be of much interest to know how such variability originates and whether it bears on the origin of new species. One species in particular, the Central American *D. pumilio,* seems most likely to yield answers to these basic questions.

D. pumilio is a small frog, usually found at elevations below 500 meters in the rain forest on the Caribbean side of Nicaragua, Costa Rica and western Panama. Variability is relatively low in Nicaragua and Costa Rica, where the frogs are usually red or reddish orange with black or bright blue hind legs. In Panama, however, variability is extremely high. It is associated with the geologic history of western Panama. Starting about 12,000 years ago the rising level of the sea created an archipelago of offshore islands and on the adjacent mainland a mixture of lowland swamp forest and rain forest. The archipelago had probably been created and reconnected to the mainland several times before. The populations of *D. pumilio* in the region show all the colors of the spectrum from red to blue; one population is even patterned solely in black and white. The different populations look like different species, and in fact we believe the complex includes at least two species. Nevertheless, most of the variation is found in what seems to be the single species *D. pumilio.*

The variation is not due simply to the independent evolution of populations isolated on islands; different parts of a single island may be inhabited by very different populations. There are also instances of intense variation within single populations. One example we have studied over a period of years is found at the northwestern end of Bastimentos Is-

Discover a small world with a big heart on the Polaris Baltic cruise.

Our Baltic/Russia cruise, stopping in Copenhagen, Stockholm, Helsinki and Leningrad will show you spectacular scenery, majestic castles and warm friendly people. And an equally engaging staff of lecturers will brief you on the region's history, culture, wildlife and heart.

The **M.S. Polaris*** offers only 1st class, outside cabins. Hearty Scandinavian food. And a chance to learn. Cruises leave from May through August. For more information call **800-223-5688** (in New York **212-751-2300**), see your travel agent, or write Dept. SA-2.

*Registered in Sweden

TO LEARN IS TO LIVE.

SALÉN LINDBLAD CRUISING
133 East 55th Street, New York, N.Y. 10022

Learn Spanish the Foreign Service way!

You can be fluent in Latin American Spanish. That's right, *fluent!* You can be speaking with all the confidence of a seasoned traveler in a matter of weeks. As well-versed in Spanish as if you worked for the U.S. Foreign Service Corps.

The Foreign Service Institute has carefully developed, tested, and refined its self-instructional audio cassette Spanish course to train the U.S. State Department's overseas personnel quickly, thoroughly, and effectively. Its clear, easy-to-follow approach and arrangement of material makes this a perfect method for successful self-instruction.

Now the FSI's complete program is available to *you.* We will send you the entire audio cassette series and instruction books to lead you step by step to fluency in weeks. You work when and where you can, at your own pace, following the taped lessons. You progress according to the schedule *you* establish.

The course uses native speakers exclusively so that you hear exactly how each word and sentence should sound. It won't be long before you start speaking and thinking in your new language. **This program has to work—our State Department depends on it!**

MONEY-BACK GUARANTEE: Try the FSI Basic or Intermediate course for three weeks. If you're not completely satisfied, return it for a full refund. **There's no risk and no obligation, so order today.**

SPECIAL OFFER: Purchase any language course from Pinnacle and get the **Toshiba KT-S3 Personal Stereo Cassette Player,** complete with headphones, FM Tuner Pack, and shoulder strap (nationally advertised for $139.00), for **only $89.95.**

French also available. Write us for details or call (212) 772-9202.

pinnacle **Dept. 219, 215 E. 79th St. New York, N.Y. 10021 (212) 772-9202**

TO ORDER BY PHONE—(Credit Card Orders Only Please)—Call Toll Free: **1-800-528-6050 Ext. 1611** In AZ 1-800-352-0458 Ext. 1611

Please send me the following course(s) shipped in attractive library albums.

☐ **Programmatic Spanish—Vol. 1 Basic:** 11 Cassettes 16 hrs. manual and 464-page text. **$100**
☐ **Programmatic Spanish—Vol. 2 Intermediate:** 8 Cassettes 12 hrs. manual and 614-page text. **$80**
☐ **Both Volumes for Only $180.** ☐ **Toshiba KT-S3. $89.95** plus $4.00 shpg.
☐ Enclosed is my check or MO for $_____ (In NY add sales tax.)
Charge my: ☐ MasterCard ☐ Visa No. _____ Expiration Date
☐ Amex ☐ Diners Club
Signature _____
Name (please print) _____
Address _____
City _____ State _____ Zip _____

land. There the color of the frogs' backs varies from orange red through bronze to metallic green. In one beachfront forest orange red frogs consistently made up about 96 percent of the population samples we took each year, with only about 4 percent of the samples showing other colors. In a forest-shaded cacao grove away from the beach, however, the orange red frogs accounted for 84 percent or less of the samples, with bronze and green frogs now constituting as much as 32 percent of the yearly totals. These consistent differences between populations of a species in different habitats are an example of "balanced polymorphism." The phenomenon is probably the genetic basis for some of the striking differences among populations of *D. pumilio.*

Although very different populations of *D. pumilio* are usually separated by stretches of forest not inhabited by the frogs, we found areas on the mainland of Panama where nearly pure populations of red frogs were in contact with similarly pure populations of green ones. Frogs of the other color were rare in both populations. A few other rare color variants in the two populations resembled frogs of more distant populations. These findings further indicate that frog populations of different colors can arise from a common gene pool.

There is more to the *D. pumilio* problem than color. The frogs in the different populations also vary in size and behavior. Their usual habitat can be on the ground or in the trees; their response to potential predators ranges from secretive and wary to bold and apparently fearless. None of these traits seems to be correlated with the frogs' defensive skin secretions, the variations in which are equally remarkable. The *D. pumilio* populations of western Panama overall produce more than 80 alkaloids representing all major classes of piperidine toxins. The number of different alkaloids secreted by frogs of individual populations ranges from six to 24, and many of these compounds have not been detected in other dendrobatid frogs. *D. pumilio* is clearly a leading candidate among all vertebrate animals for the title "most variable species," but we are a long way from understanding all the factors that have made it so.

EIGHT FROGS representing different populations of *Dendrobates pumilio* are shown in dorsal and ventral position in the painting on the opposite page. These samples show only part of the extraordinary divergence among frogs of the *D. pumilio* population complex on the Caribbean side of western Panama. All the frogs are shown at the same size, but in life they vary in size as well as in color; they are also remarkably diversified in behavior and in the alkaloids of their defensive secretions.

How to stay on top of your inventory.

To help give profits a lift, a businessperson could use the IBM Personal Computer. With Inventory Control by Peachtree Software Inc., you can stay on top of it all.

Keep track of each item in stock. Maintain vendor and cost files. And record period-to-date and year-to-date sales amounts and merchandise quantities.

The IBM Personal Computer will also help monitor sales performance and profit margins. Help you decide when and how much to reorder. Even print your management reports.

To take stock of the IBM Personal Computer, visit an authorized dealer soon. You'll see that the quality is what you'd expect from IBM. The price isn't.

IBM

The IBM Personal Computer
A tool for modern times

For more information on where to buy the IBM Personal Computer, call **800-447-4700.** In Alaska or Hawaii, **800-447-0890.**

Fan Vaulting

In the 14th century English masons converted the pointed Gothic arch into an architectural shell of flowing curvature. Their aims were primarily aesthetic, but the vaults are soundly engineered

by Walter C. Leedy, Jr.

One of the central problems in architecture is how to enclose and articulate interior space. The problem becomes more critical as the size of the space increases, and it was particularly acute in a medieval cathedral or abbey. There the upward sweep of the walls strongly emphasized the ceiling, which was consequently expected to have a striking design. Furthermore, the ceiling was generally made of stone and was very heavy. The solution adopted in the most representative Gothic churches was to construct the ceiling as a series of vaults in which planar sheets made up of stone blocks ran between sharply pointed arches. The arches served as ribs bearing at least part of the load.

As medieval masons gained experience with ribbed vaults they found that construction could be made easier and the design more versatile by including additional ribs in the ceiling structure. Eventually the vaulting came to be a complicated lacework of ribs, which could be given many shapes. In England in the 14th century the proliferation of ribs led to the invention of an entirely new structure: the fan vault. In the fan vault the planar stone sheets of the ribbed vault were replaced by a rounded structure called a conoid that resembles a cone cut in half along its axis and erected with the pointed end down. Four conoids were joined to span each modular unit of the vaulting. On the curved surface of each conoid numerous ribs radiate upward and outward toward the peak of the ceiling. It is from this fanlike pattern of ribs that the style takes its name. Between the ribs is intricately carved stone tracery.

In a ribbed vault the arches were expected to bear the weight, but the conoid of a fan vault can act as a shell structure, in which the stresses are distributed fairly evenly throughout the stone fabric. The medieval builder had no way to analyze the stresses in a shell. Nevertheless, modern engineers have shown that the conoid of a fan vault is a very stable structure. Between about 1350 and 1540 English designers exploited this quality to construct more than 100 fan-vaulted ceilings spanning increasingly large spaces. Among the most famous are the ceiling of the chapel of King's College at the University of Cambridge and that of the Chapel of Henry VII in Westminster Abbey.

The elaborate ceiling of a fan-vaulted building provides a powerful visual experience. The rounded conoids flow into one another along the planes of the wall and give an impression of gracefully and continuously molded volumes of space, in contrast to the angularity of the ribbed vault. The intricate carving on the surface of the conoid emphasizes its volume and contributes to the impression of unity. The visual effect was probably foremost in the minds of the masons who built the fan vaults. It should not be forgotten, however, that underlying the aesthetic experience is the solution to a fundamental architectural problem arrived at by designers relying on practical experience in the absence of highly refined theoretical skills.

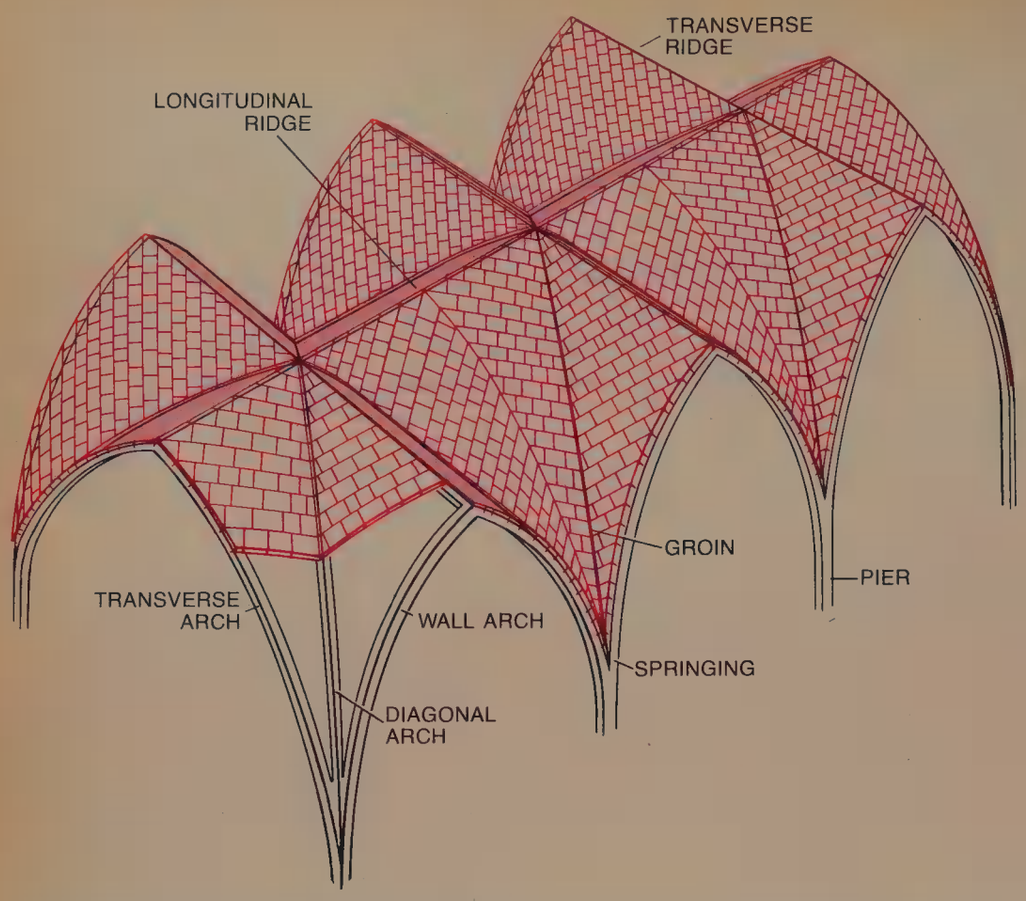
The development of the fan vault is closely connected to the physical properties of stone and the aesthetic preferences of the English masons and their patrons. Stone can withstand large compressive forces, but it has little tensile strength: it cannot readily withstand being pulled or stretched. The downward pull of gravity can cause substantial tensile stress in a horizontal beam supported at both ends. In a stone ceiling the tensile stress must somehow be converted into a compressive stress. In the pointed arch of the ribbed vault the conversion is accomplished by raising the arch stones above the horizontal plane defined by the points from which the arch springs. The force of gravity is thereby transformed into outward and downward compressive forces that are transmitted through the stones of the arch toward the walls.

The modular unit of a vaulted ceiling is usually a rectangular compartment called a bay, whose long axis is perpendicular to the major axis of the hall. In a ribbed vault the bay generally includes three pairs of arches. The transverse arches cross the hall. The wall arches run along the wall at the short sides of the rectangle, most often with large clerestory windows in the wall under the arches. The opposite corners of the bay are connected by diagonal arches that intersect in the center of the rectangle. Tall piers placed at the springing, the point at the corner of the bay where the lower segments of three arches meet, help to support the vaulting high above the floor. The thrust from the arches is conveyed to the ground by the piers and by flying buttresses that meet the exterior of the wall near the springing.

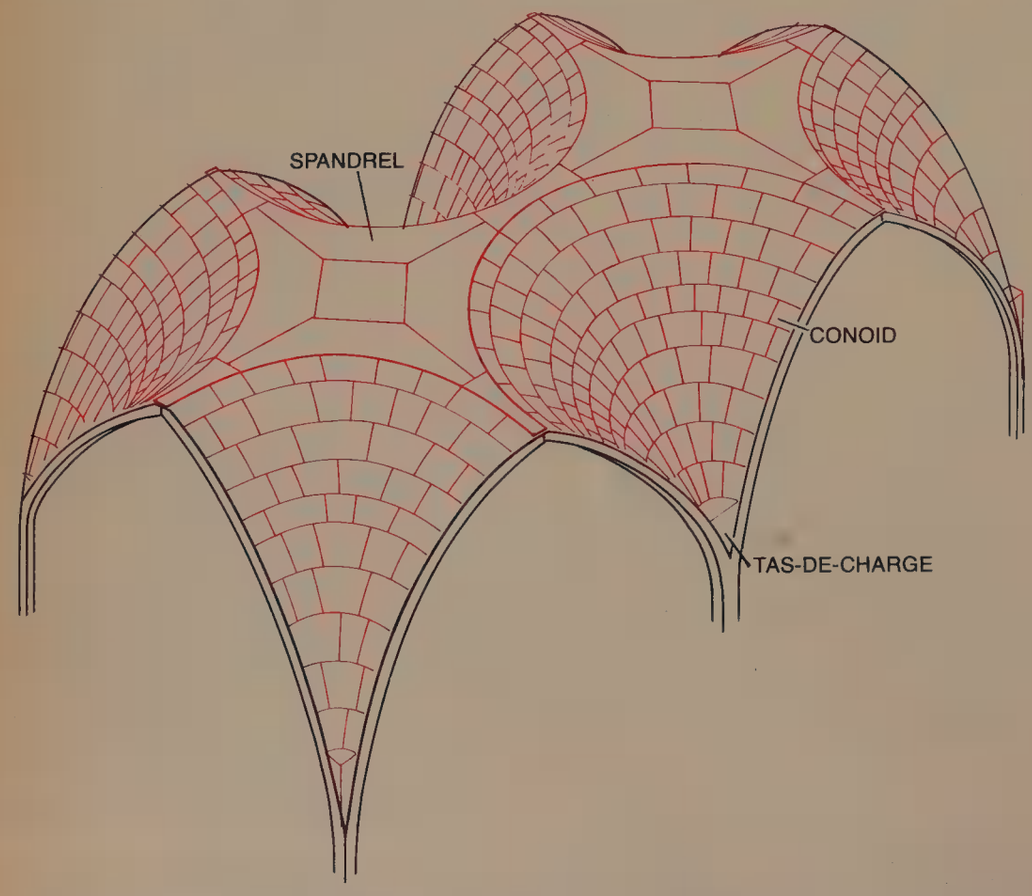
In the construction of a rib-vaulted ceiling the arches were erected first, utilizing a complex wood scaffolding called centering as a form on which to lay the stones. When the six arches in each bay had been erected, they served as the framework for the web: the sheets of stone laid between the ribs.

Ribbed vaulting, such as that in the cathedrals of Amiens and Rouen, has a strongly linear appearance, which is the basis of Gothic artistry. The effect was achieved by emphasizing the vaulting ribs as a series of frames with the web filling the space between them. Further-

KING'S COLLEGE CHAPEL at the University of Cambridge has one of the most beautiful examples of fan vaulting. The main vault of the chapel, shown here, was completed in 1515. The ceiling has the longest span of any known fan vaulting: the distance across the hall is 12.7 meters. The rounded intersecting surfaces are the main weight-bearing structures of the fan vault. Called conoids, they act as shells in which stresses are distributed fairly evenly. The carving, derived from the tracery of stained-glass windows, serves to unify the building's interior.



RIBBED VAULT, the classic Gothic ceiling design, was conceived by medieval masons as a frame structure in which pointed arches serve as weight-bearing ribs. Each of the modular units of the vaulting, called a bay, includes six arches arranged on a rectangle with intersecting diagonals. The load of the ceiling is converted into outward and downward thrusts that are conveyed in part through the arches to the piers and walls. The surface of the vault between the arches is filled with stone sheets called webs. In England the stones of the web were laid at an angle of approximately 45 degrees from the longitudinal or transverse ridge to the groin.



FAN VAULT developed from the ribbed vault in the 1350's. The fan-vault conoid is a surface of rotation, created by rotating a curve about a vertical axis. Viewed from the center of a bay, each conoid has a convex horizontal section and a concave vertical section. This form allows the conoid to act as a set of vertical and horizontal arches that convey thrust to the wall.

more, the large transverse arches draw attention to the division of the ceiling into bays. In contrast, although the fan vault is still composed of linear elements, the unity of the interior space is emphasized. The viewer's gaze is drawn along the ceiling by the repetitive and graceful form of the conoids; the distinction between bays generally has little visual importance.

Although the aesthetic principles of the two kinds of vaulting are quite different, the historical transition between them was not abrupt. On the contrary, the fan vault developed gradually out of the construction of ribbed vaults. Ribbed vaulting was common in England, France and other countries of continental Europe, whereas fan vaulting developed only in England; the question naturally arises of why the new form emerged only there.

The answer lies in the aesthetic values and building practices that prevailed in England in the Middle Ages. The comparison with France is instructive. In France the courses, or rows of stone, in the web were laid parallel to the sides of the rectangle covered by the vault, so that the courses run in horizontal rows from the bottom of the vault to the top. Each quadrant of the bay resembles a peaked roof; the ridge at the peak of the quadrant is called a transverse or a longitudinal ridge depending on the orientation of the quadrant. Along the groin, where the quadrants meet, the courses of the web were laid parallel to both ridges, forming right angles to the sides of the bay. For a while English masons followed the French precedent. Soon, however, they began to lay the courses from the ridge of the quadrant to the groin at an angle of about 45 degrees to both ridges. This style of coursing led in part to the conoid's being conceived of as a rounded structure.

English practice came to differ in other ways as well. In the French vault the web courses were laid over the top of the rib; the rib and the web were connected only by mortar. In England the upper edges of the rib were deeply notched and the web stones were set into the notches before the stones were cemented together. This practice, called rebating (or in modern English rabbeting), tended to make the rib and web a single assembly. Here is architectural evidence that the English mason saw the rib as a weight-bearing member.

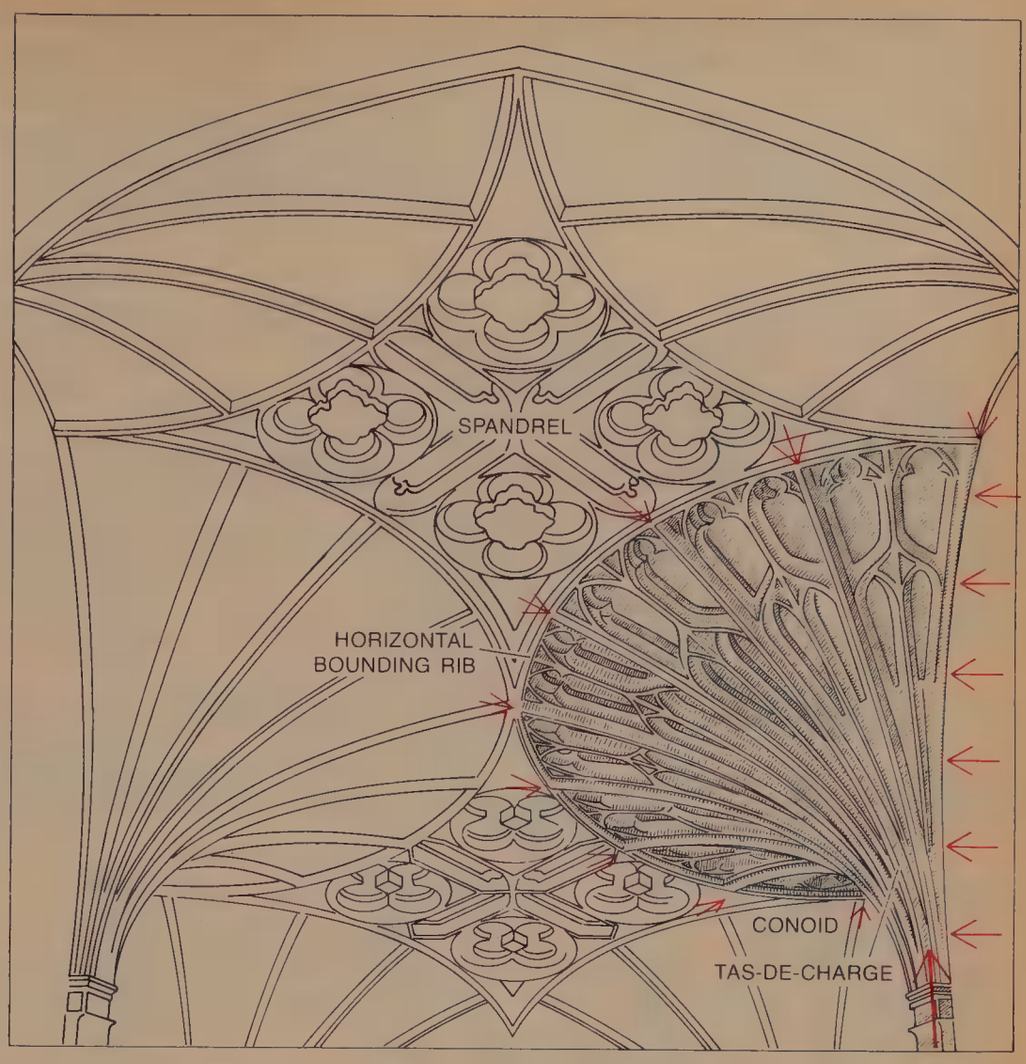
Even the design of the arches in England differed from that in France. In France height, illumination and openness of structure were the most highly prized architectural qualities. To increase the area of the clerestory windows the wall arches were often stilted by raising the bottom of the arch above the springing on vertical wall shafts. Stilting had several significant conse-

quences. The web near the window had to be warped to accommodate the shape of the arch. In addition the area of the wall between the clerestory windows was much reduced and the outward thrust from the vaulting was concentrated on a narrow strip of wall. As a result the flying buttresses had to be placed near their optimal position or the wall would collapse.

In England there was less concern with height and openness. The walls were generally lower and thicker. The clerestory windows were smaller and hence the wall arches did not have to be stilted. This increased the area of the wall across which the vaulting thrust was distributed and reduced the need for buttressing.

The ribbed vault spread through much of Europe in the 12th century. Medieval masons soon learned that the addition of two further types of ribs could increase the flexibility of the vaulting design. The first type, the tierceron rib, ran from the springing to the ridge of the vault, roughly bisecting the angle between the diagonal rib and either the transverse rib or the wall rib. The lierne rib was a shorter structural member inserted at an angle between the longer ribs.

English masons built numerous tierceron vaults. The tierceron design was well suited to the English style, in part because it made possible great flexibility in adjusting the curvature of the ribs. In contrast to architectural practice in France, where rib curvature was constrained by the striving for height and illumination, in England the tierceron ribs radiating from the springing could be formed into an appealing geometric pattern.



CONOID OF A FAN VAULT must be compressed along all its edges to be in equilibrium. The drawing shows the vaulting in the Church of St. Mary at North Leigh, Oxfordshire, which was finished in about 1440. The heavy spandrel panel between the conoids provides a compressive force along the horizontal bounding rib and serves as the keystone for the entire vault. The tas-de-charge built out from the wall at the corner of the bay supports the bottom of the conoid. The walls provide inward thrust to oppose the thrust transmitted outward through the conoids. Ribs carved on the surface are primarily aesthetic rather than structural elements.

In the high Gothic ribbed vault the volume defined by the webs intersecting along the groin in the corner of the bay was quite irregular and the shape was determined by other design considerations. As more ribs were added to the tierceron vault the volume became more regular and began to appear as a satisfying form worth designing for its own sake. At the same time the English coursing pattern began to yield an intriguing result. As I have mentioned, in England the web courses were laid at an angle from the longitudinal or transverse ridge to the groin. In tierceron vaulting the web courses in the corner of the bay began to approximate the form of regular, horizontal polygons. The polygons were concentric on the springing of the vault and increased in size toward the peak of the ceiling.

By the early part of the 14th century the main preconditions for the fan vault had been created in tierceron vaulting with concentric courses and rebated ribs. Two further technical advances were necessary for the fan vault to come into being. The first was the practice of carving a segment of the rib and the adjoining portion of the web from the same block of stone. The carved stones were then fitted together to form the vaulting much as blocks of ice are fitted to form an igloo. This kind of construction, which is termed jointed masonry, was employed in much fan vaulting.

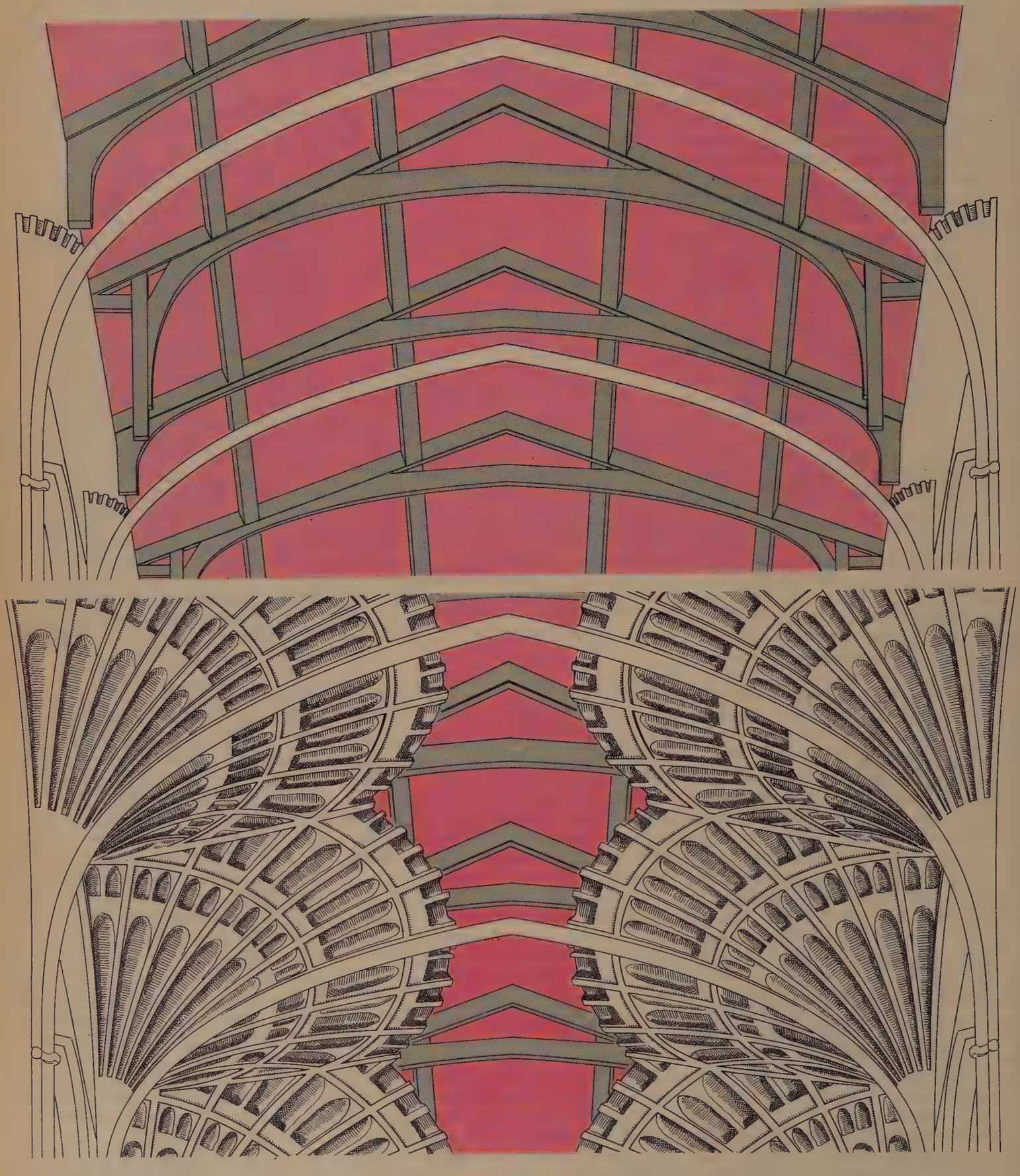
The use of jointed masonry by English masons does not seem to have been motivated primarily by engineering considerations. On the contrary, it seems to have been largely the result of aesthetic concerns. English builders valued a high degree of carved articulation of the vaulting surface. Such articulation is much easier to achieve with large, closely fitted blocks than with ribs and small panels of stone inserted between them. As the size of the blocks was increased to provide a carving surface, the blocks tended naturally to assume a structural function.

The introduction of jointed masonry led to another innovation: the major axis of the rectangular cross section of all the ribs was made perpendicular to the vaulting surface. In a ribbed vault the major axis was generally perpendicular to the floor. When jointed masonry is used in a fan vault, however, the joint between the portions of a rib on two different blocks can be made perfect only if the ribs are perpendicular to the vaulting surface. The proliferation of ribs and hence of intersections between ribs made this an increasingly important aesthetic consideration.

By the middle of the 14th century the major elements of the fan vault had evolved in England. According to most modern investigators, the defining characteristics of a fan vault are as follows: vaulting conoids with a regular geometric form; regularly spaced ribs, all with the same curvature; a distinct spandrel, or central ceiling panel; ribs perpendicular to the surface of the vaulting, and patterned surface carving. This set of properties seems to have appeared first in small square canopies in the tombs of great nobles. (The origin of the style may never be identified with certainty because of the destruction of many im-

portant sites during the reign of Henry VIII.) Because the tomb vaults are small, they are often referred to as toy, or decorative, fan vaults. The first structural fan vaulting could well be that in the Trinity Chapel of Tewkesbury Abbey in Gloucestershire; the vaulting was completed in about 1380. The first large structural fan vaulting and one of the most significant early examples of the form is the vaulting in the cloister of Gloucester Cathedral, which was begun in the second half of the 14th century.

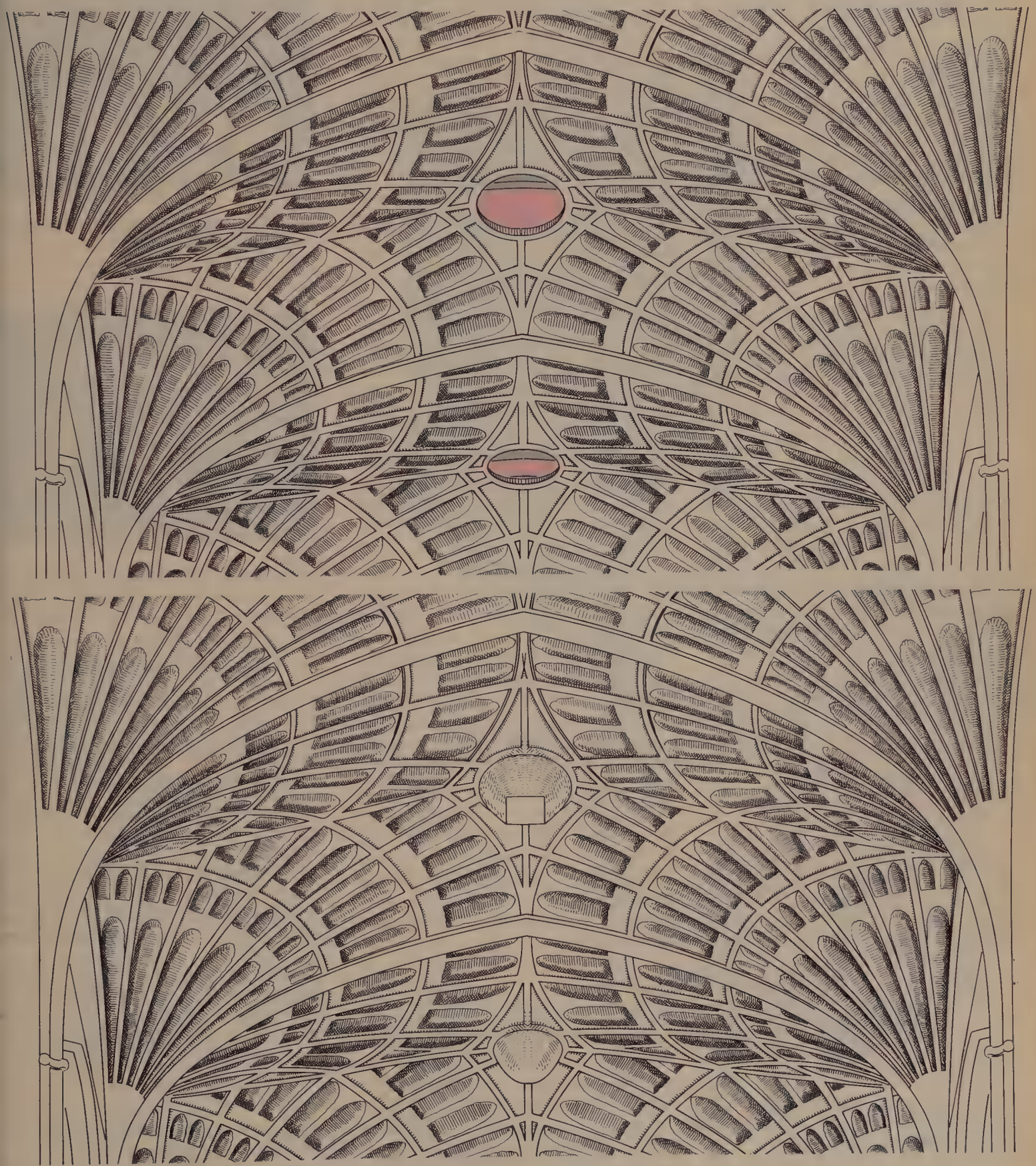
An essential feature of the fan vault is that the vaulting conoid is a surface of rotation, that is, a form produced by rotating a curved line about a vertical axis. The axis is the corner of the bay and the curved line is the rib. One consequence of this form is that the conoid has a horizontal section that is a circle or a portion



CONSTRUCTION OF VAULTING in King's College Chapel was accomplished in four stages. The stone blocks were carved in workshops next to the building site. It seems likely they were put in place with their decorative surfaces only partially finished. Wood centering, or scaffolding, was utilized as a form for the vaulting during construction. After the walls and roof were up the large transverse arches that divide the ceiling into bays were erected (*upper left*). The conoids were then laid, proceeding inward from the transverse arch-

of a circle at every vertical level. The circle is convex with respect to the center of the bay. The rib, however, is concave with respect to the center of the bay. Thus at every point on the surface of the conoid a convex horizontal curve intersects a concave vertical curve.

Since the conoid extending from the corner of the bay is circular, fan vaulting is best suited to a square bay. (Fan vaults were also built in rectangular modules, but ingenuity was required to adapt the form to the bay.) In a square bay the conoid has a horizontal section that is a quarter of a circle. At the top of the ceiling is an approximately diamond-shaped space formed by the intersection of the four circular conoids. This space is filled by the spandrel. At the curved boundary between the spandrel and the conoid the builder often inserted



toward the center of the bay (*lower left*). The conoids were assembled one horizontal level at a time. When the horizontal arches of adjoining conoids met in the center of the bay, the large transverse ridge stone that serves as the keystone for both arches was put in position. The stones of the longitudinal ridge were then laid (*upper right*). Finally, the large, heavy boss at the center of each spandrel was dropped in place from above (*lower right*). After the centering was taken down any final carving of the ribs and tracery was done.

Physician, did you miss any of these significant developments in medical science?

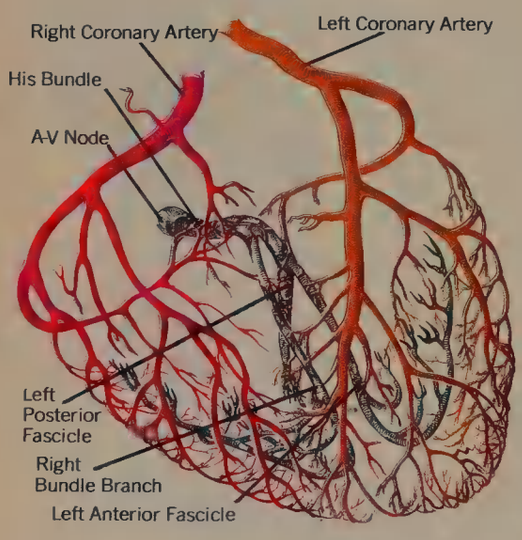
- ***Campylobacter fetus* subsp. *jejuni* is associated with a colitis that can clinically and sigmoidoscopically resemble acute idiopathic ulcerative colitis. Stool cultures are in order for *C. fetus* before beginning nonspecific anti-inflammatory therapy.**

- **Coumarin derivatives cross the placenta. A recent study shows that the consequences for the fetus can be severe. These include embryopathy, stillbirth, and premature delivery.**

- **Nonsteroidal anti-inflammatory drugs may produce a marked reduction in glomerular filtration rate; with termination of the drug, GFR returns to normal.**

- **The first documented incident of indigenous transmission of dengue in the continental United States since 1945 has been reported in Brownsville, Texas.**

IF THESE ITEMS are familiar you must be a prodigiously energetic or prodigiously lucky reader. With 2,000 or more journals published each year, information that significantly affects patient management all too easily slips by. Textbooks are out-of-date before they are published.



Branches of the right and left coronary arteries supply blood to the A-V node and intraventricular conduction system.

SCIENTIFIC AMERICAN *Medicine* is lucidly illustrated with drawings and photographs. Some examples are seen here and on the facing page.

SCIENTIFIC AMERICAN *Medicine* is the busy clinician's answer to this problem.

Because its authors update SCIENTIFIC AMERICAN *Medicine* every month, it is always current. Because the new information appears in a single source, it is there when you need it.

This 2,500-page, innovative union of publishing and electronic technology is the work of leading scholar-practitioners from Harvard and Stanford. The editors are Edward Rubenstein, M.D., F.A.C.P., and Daniel Federman, M.D., F.A.C.P.

Each month as authors update their contributions, revisions are entered on the magnetic tape on which the text and index are stored. The tape drives high-speed phototypesetting equipment so that subscribers receive about eight new chapters and a new index every four weeks; a bulletin highlights new developments.

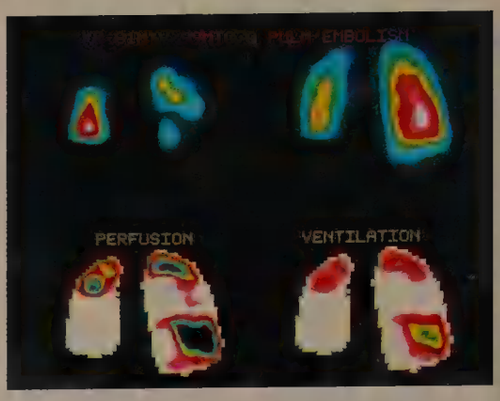
New material replaces old material in the living text, so that the information is there — up-to-date, at your fingertips.

A CME program of eight patient management problems offered over a 12-month period is available at no extra cost. As an organization accredited for continuing medical education, the Stanford University School of Medicine designates this continuing medical education activity as meeting the criteria for 32 credit hours in Category 1 for Educational Materials for the Physician's Recognition Award of the American Medical Association, provided it has been completed according to instructions. This program has been reviewed and is acceptable for 32 prescribed hours by the American Academy of Family Physicians. This program has been approved by the American College of Emergency Physicians for 32 hours of ACEP Category 1 credit.

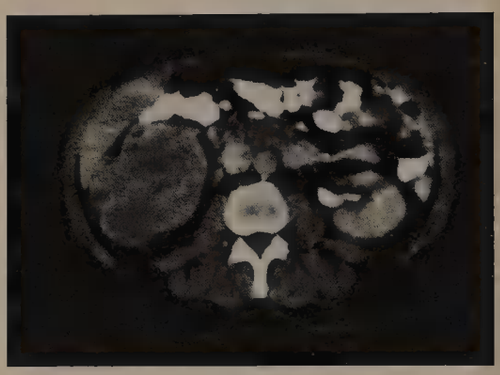
Trial Offer

We invite you to try SCIENTIFIC AMERICAN *Medicine* — for two months at no cost. Send us the coupon and you will receive the two-volume text and two monthly updates. You may also take a CME test for credit. At the end of 60 days, if you decide to continue the subscription, we will bill you for $220 for the full 12 months (renewal is currently $170); otherwise return the two volumes.

Please mail the coupon today and let us take the hassle out of keeping up.



Computerized scintigraphy reveals pulmonary thromboembolism.



Abdominal computed tomogram reveals large renal carcinoma replacing part of right kidney.

THE DISTINGUISHED AUTHORS AND THE FIFTEEN SECTIONS OF SCIENTIFIC AMERICAN ***MEDICINE***

1. Cardiovascular Medicine

Edgar Haber, M.D., Harvard Medical School and Massachusetts General Hospital
E. William Hancock, M.D., F.A.C.P., Stanford University School of Medicine
Roman W. DeSanctis, M.D., Harvard Medical School and Massachusetts General Hospital
Adolph M. Hutter, Jr., M.D., F.A.C.P., Harvard Medical School and Massachusetts General Hospital
Eve Elizabeth Slater, M.D., Harvard Medical School and Massachusetts General Hospital

2. Dermatology

Eugene M. Farber, M.D., Stanford University School of Medicine
Elizabeth A. Abel, M.D., Stanford University School of Medicine

3. Endocrinology

Daniel D. Federman, M.D., F.A.C.P., Harvard Medical School

4. Gastroenterology

Gary M. Gray, M.D., Stanford University School of Medicine

Peter B. Gregory, M.D., Stanford University School of Medicine

John Austin Collins, M.D., Stanford University School of Medicine

Douglas Wilmore, M.D., Harvard Medical School and Brigham and Women's Hospital

5. Hematology

Stanley L. Schrier, M.D., Stanford University School of Medicine

6. Immunology

John David, M.D., Harvard Medical School and Harvard School of Public Health

7. Infectious Disease

Thomas C. Merigan, M.D., Stanford University School of Medicine

Morton N. Swartz, M.D., F.A.C.P., Harvard Medical School and Massachusetts General Hospital

Cyrus C. Hopkins, M.D., Harvard Medical School and Massachusetts General Hospital

Adolf W. Karchmer, M.D., F.A.C.P., Harvard Medical School and Massachusetts General Hospital

Robert H. Rubin, M.D., F.A.C.P., Harvard Medical School and Massachusetts General Hospital

Harvey B. Simon, M.D., F.A.C.P., Harvard Medical School and Massachusetts General Hospital

Peter F. Weller, M.D., Harvard Medical School; Beth Israel Hospital, Boston; and Brigham and Women's Hospital

8. Intensive and Emergency Care

Edward Rubenstein, M.D., F.A.C.P., Stanford University School of Medicine

9. Metabolism

George F. Cahill, Jr., M.D., Harvard Medical School, Howard Hughes Medical Institute; and Brigham and Women's Hospital

10. Nephrology

Roy H. Maffly, M.D., Stanford University School of Medicine and Palo Alto Veterans Administration Medical Center

William Bennett, M.D., Health Sciences University, Portland, Oregon

11. Neurology

Robert W. P. Cutler, M.D., Stanford University School of Medicine

12. Oncology

Saul A. Rosenberg, M.D., F.A.C.P., Stanford University School of Medicine

13. Psychiatry

Ned H. Cassem, M.D., Harvard Medical School and Massachusetts General Hospital

14. Respiratory Medicine

Eugene D. Robin, M.D., F.A.C.P., Stanford University School of Medicine

15. Rheumatology

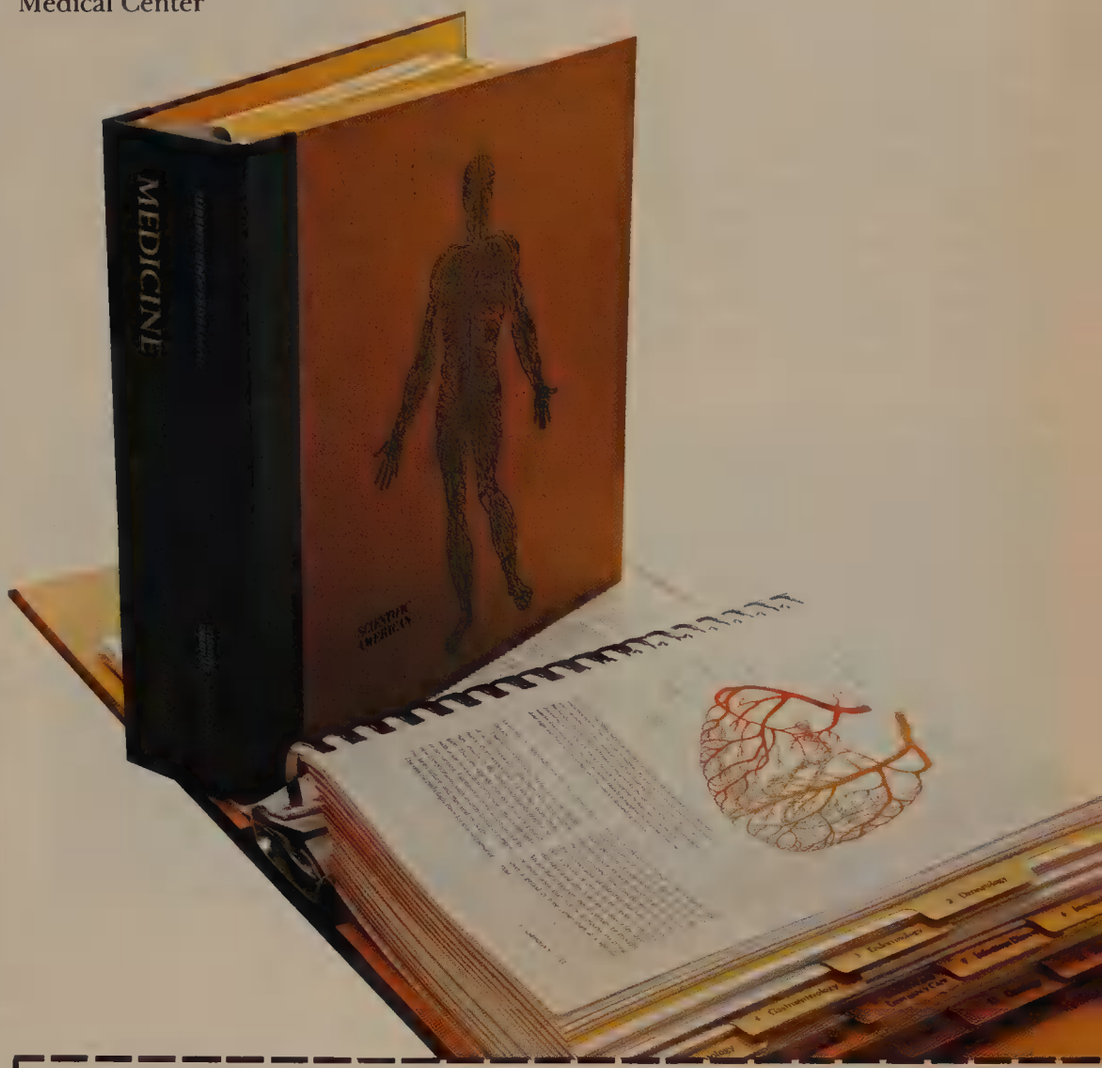
Stephen M. Krane, M.D., Harvard Medical School and Massachusetts General Hospital

Dwight R. Robinson, M.D., Harvard Medical School and Massachusetts General Hospital

Andrei Calin, M.D., M.A., M.R.C.P., Stanford University School of Medicine and Palo Alto Veterans Administration Medical Center

Order by Phone

You can order SCIENTIFIC AMERICAN *Medicine* by telephone. **Please call this toll-free number: 1-800-345-8112 (in Pennsylvania, call 1-800-662-2444);** you will be billed after your subscription begins. Toll-free calls are acceptable only for orders placed in the continental United States.



SCIENTIFIC AMERICAN MEDICINE

415 Madison Avenue, New York, N.Y. 10017

Please enroll me as a subscriber to SCIENTIFIC AMERICAN *Medicine*. On receipt of this coupon you will send me the advanced two-volume text described in your announcement and update it regularly by sending me new monthly subsections. I understand that the price of $220 for first year of service is tax deductible, as is current renewal at $170. If I am not entirely satisfied, I may cancel at any time during the first 60 days, returning all materials for a *complete refund.*

☐ Please enter my subscription for SCIENTIFIC AMERICAN *Medicine*
☐ I shall also enroll in the CME program
☐ I enclose a check made out to SCIENTIFIC AMERICAN *Medicine* for $220*
☐ Bill me ☐ VISA ☐ MasterCard

Expiration Date_____ Account Number_____
*Please add sales tax for California, Illinois, Massachusetts, Michigan and New York

Name_____

MD Specialty_____

Address_____

City_____ State_____ Zip_____

Signature_____

Please allow six to eight weeks for delivery. All payments must be in U.S. dollars. CME is available outside the U.S. and its possessions for an additional $65. Surface routes used unless airmail delivery is prepaid.

a horizontal bounding rib to divide the two structures visually.

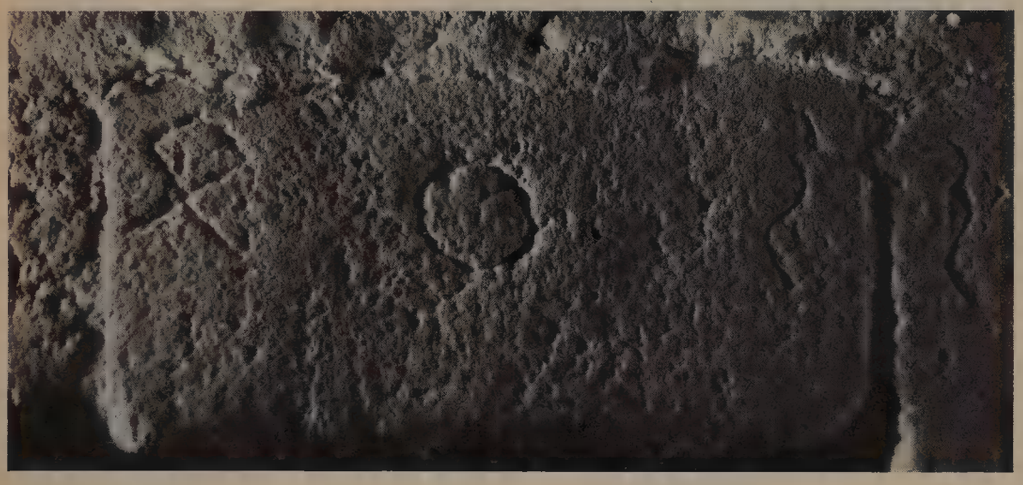
The medieval mason was no formalist. He felt free to alter the elements of the fan vault in order to accommodate the design requirements or his own taste. For example, in the chapel of King's College the vertical ribs appear to be equally spaced, but actually there are small discrepancies resulting from the fact that the decision to fan-vault the ceiling was made after the building had been partially erected. Sherborne Abbey in Dorsetshire has beautiful examples of fan vaulting that were quite influential among 15th-century designers. The conoids in the chancel (the area that includes the altar) have horizontal sections that are polygonal rather than circular. Moreover, the conoids are constructed of separate ribs and small flat panels rather than jointed masonry. Nevertheless, in both King's College and Sherborne Abbey the ceilings must be described as fan vaulting because of their overall appearance, in spite of the deviations from the definition of an idealized fan vault.

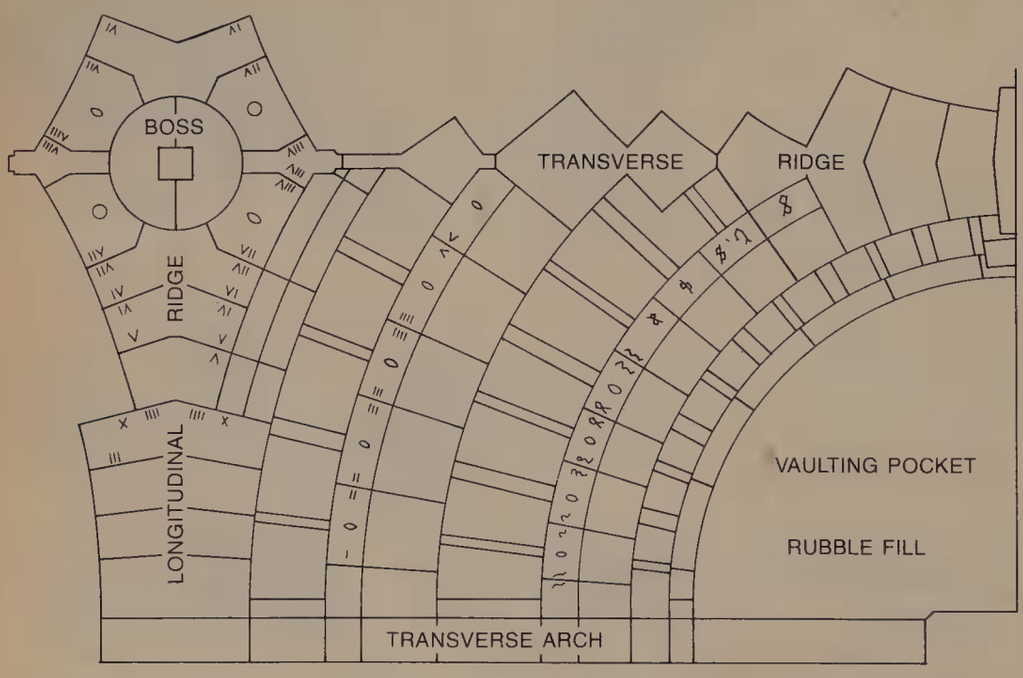
The construction of fan vaults can be divided into three periods beginning with the construction of toy fan vaults in tomb canopies. The first period ended with the completion of the cloister of Gloucester Cathedral in 1412. Between 1412 and 1430 no major fan vaults were built. This was a time of labor shortage, high taxes and economic depression; fan vaulting was costly and it is likely that no individual or institution could muster the capital needed to undertake construction. The design of the Sherborne Abbey chancel vaulting in the late 1430's marks the beginning of the second period, which lasted until about 1475. In the third period, from 1475 to about 1540, many of the largest and most important fan-vaulted ceilings were constructed.

The overwhelming majority of fan vaults were built in ecclesiastical buildings, and a considerable number of these were constructed in chantry chapels. A chantry was a fund established to pay for masses for the soul of the founder; the chantry chapel was where the masses were said. Most of the early chantry chapels were endowed by noblemen and were intended as much to call attention to the greatness of the founder as to ensure the safety of his soul. The fan vault with its striking visual form and intricately carved tracery was well suited to the purpose.

In the second half of the 15th century the taste for fan vaulting spread to the middle class and to the King. St. George's Chapel at Windsor Castle, which was a royal institution, has fan vaults dating from the 1480's. Shortly after 1500 work began on the most magnificent of all fan vaulting: the ceilings in the Chapel of Henry VII at Westminster Abbey, which Henry conceived as a huge chantry chapel for the Tudor dynasty. Hence the church, the nobility, the middle class and the crown all contributed to the diffusion of the new architectural form.

The spread of fan vaulting was accompanied by an increase in the scale of the projects. The vaulting completed about 1380 in Trinity Chapel at Tewkesbury Abbey spans 1.7 meters across the hall. The main vault of the King's College chapel, which was completed in 1515, spans 12.7 meters. This is the longest span of any fan vaulting. One reason it took English builders more than 150 years to increase the span of the vaulting to its maximum is that they had little conception of how a fan vault works.

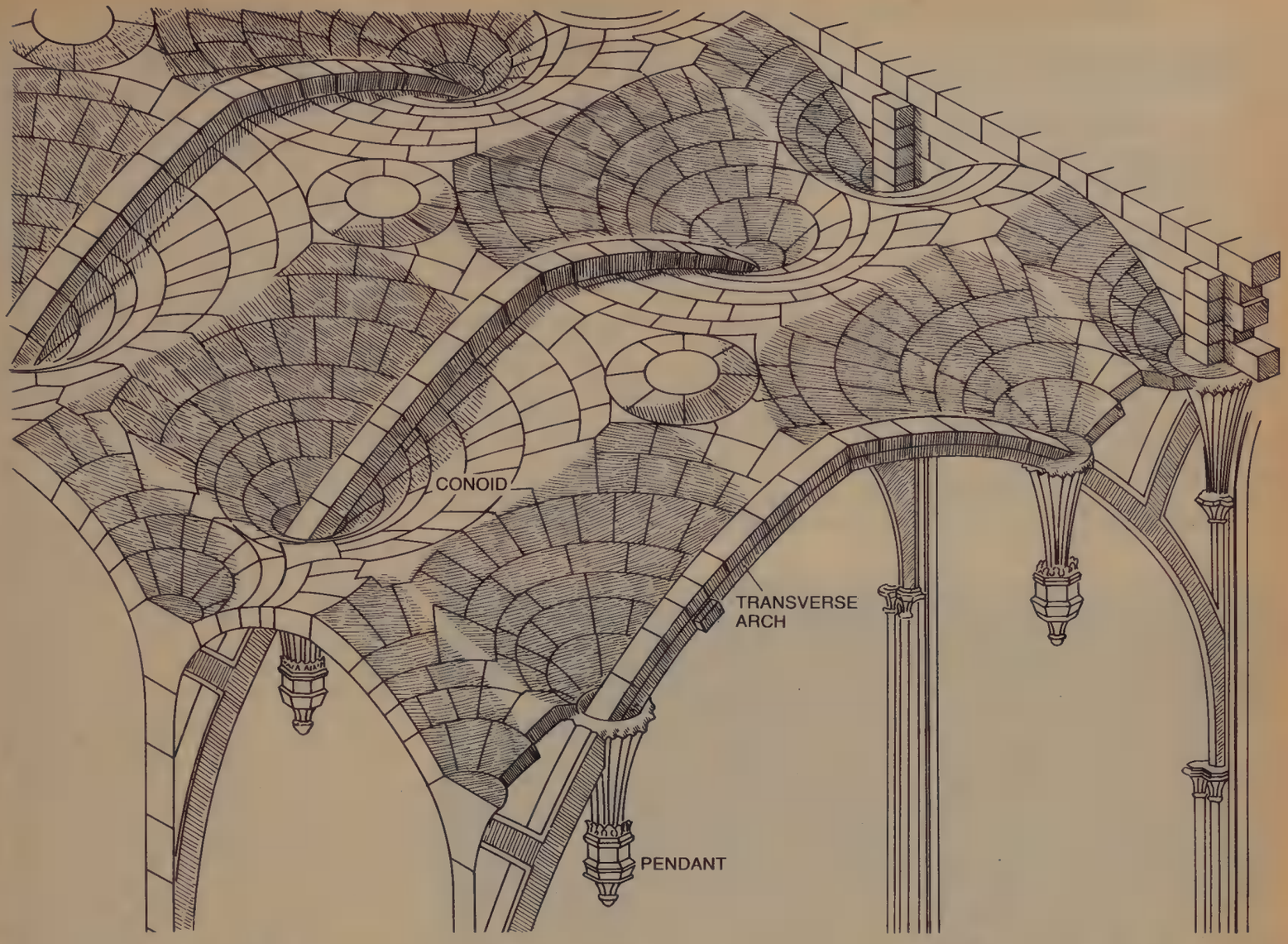




PLACEMENT MARKS on the stones for the vaulting of King's College Chapel enabled the stone setters on the building site to complete the ceiling by merely following the marks. The symbols were carved in the workshops where the vaulting was prefabricated. The large circular mark at the center of the stone indicates the quadrant of the bay in which the stone was placed. The marks at the ends of the stone next to the joints are the Arabic numerals 3 and 4. They indicate the horizontal level in the conoid where the stone was to go, the position in that level and the timing of insertion. The placement marks can still be seen on the upper surface of the vaulting stones, as is indicated by the plan view of one quadrant of a bay in the main vault.

Twentieth-century engineers utilizing advanced mathematical tools have shown that for the conoid of a fan vault to be in equilibrium it must be supported along all its edges. A substantial weight must apply a compressive load at the upper edge along the horizontal bounding rib that separates the conoid from the spandrel. The load is provided by the spandrel itself, which is a very heavy stone plate. The large bosses, or raised decorative stones, that constitute the center of the spandrels in the King's College chapel weigh some 1,400 kilograms each.

The spandrel serves as the keystone of the vertical arches in the conoid. As in the ribbed arch, the thrusts are directed outward and downward. The downward thrusts are conveyed to the bottom of the conoid, where the structure rests



CHAPEL OF HENRY VII in Westminster Abbey has the most magnificent of all fan vaulting. Work on the vaulting was begun in about 1500. The ingenious design of the ceiling combines frame structures and shell structures. Substantial transverse arches support large pendants near the wall of the bay. The conoids are built over the pendants. At the pendants the transverse arches run up through the vaulting and along its upper surface, so that the arches are not visible in the central part of the ceiling. Most of the weight is taken up by the arch where it passes through the conoid at the pendant. The arch functions as a frame member and the conoid as a shell structure.

against the wall. At this point the conoid is supported by the tas-de-charge, a projection built out from the wall in the corner of the bay. As noted above, the conoid is composed of concentric courses radiating upward from the springing, with each course having the form of an arch whose alignment is approximately horizontal. The outward thrust is conveyed through the horizontal arches to the wall, which provides an opposing thrust. Thus the conoid is compressed between the spandrel, the walls and the tas-de-charge.

The stresses in the conoid tend to be distributed fairly evenly rather than concentrated in the ribs. The specific way the stresses are distributed has important implications for the stability of the conoid, which, as I have noted, is a shell structure. An architectural shell is defined as a stressed structure much thinner than it is broad. The thrust acting on each point of the shell can be represented by a vector with components directed outward and downward; the set of all such vectors forms an imaginary surface called the thrust surface. Mathematical analysis shows that for the conoid to be in equilibrium the thrust surface must lie within the physical surfaces of the conoid shell. If the thrust surface emerges from the surfaces of the conoid, the structure is likely to collapse.

The builders of fan vaults had no way of knowing this fundamental fact. The analysis of stresses in a shell was far beyond their theoretical capacities. On the basis of long experience, however, the masons developed several techniques that greatly increased the stability of the vaulting. The narrow volume at the base of the conoid above the tas-de-charge is termed the vaulting pocket. In many fan-vaulted ceilings the pocket was filled to a height of about a meter with a solid rubble made up of stone chips and cement. In the chapel of King's College the rubble reaches precisely the same height in each conoid, indicating that it was a planned element of the structure rather than a casual afterthought.

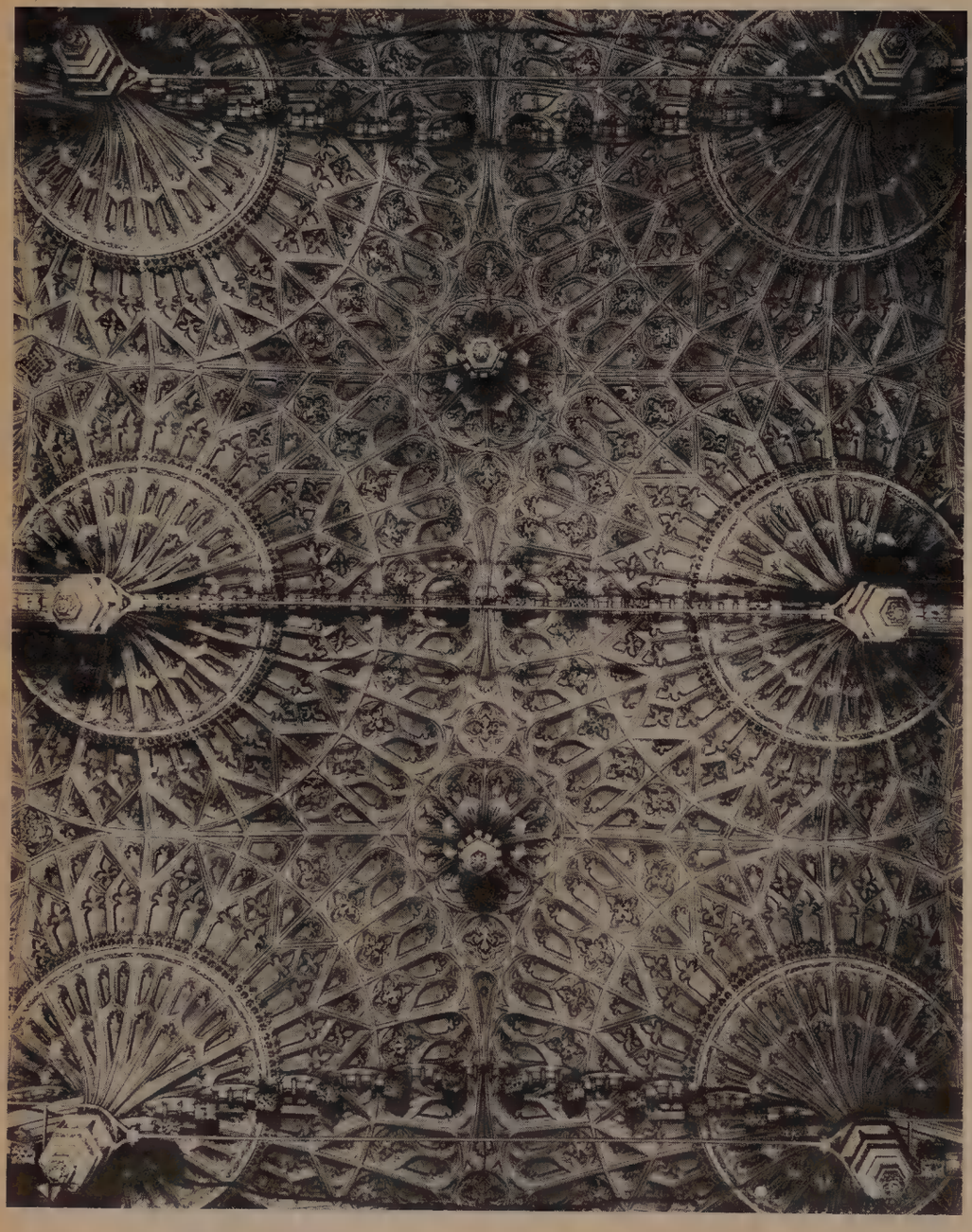
The presence of the rubble fill has three significant consequences. First, the thrust surface extends into the rubble fill, which conveys the thrust to the wall. In this way the thrust is distributed over a much larger area of the wall than it would be if the shell alone transmitted it to the wall. Second, the rubble acts as a weight cantilevered over the floor. The weight opposes the outward thrust of the conoid reaching the walls in the corner of the bay and thereby further reduces the stress put on the walls.

Most significant for the stability of the conoid, the rubble fill reduces the span of the vaulting that functions as a shell structure. The rubble cannot act as a shell because it is a substantial three-dimensional form rather than a thin sheet. Since the vaulting thrust passes into the rubble, the lower part of the conoid does not act as a shell. Jacques Heyman of the University of Cambridge has shown that the part of the vaulting that func-

tions as a shell is the part between the top of the rubble fill and the lower edge of the spandrel. This distance can be quite short in relation to the entire span of the vaulting. Reducing the length of the shell increases the ratio of thickness to length. The shorter and thicker the shell, the greater the probability that the thrust surface lies between the surfaces of the conoid.

Thus without significant theoretical knowledge the mason managed to make the fan vault quite stable. If stability can be ensured, the only other major risk to be considered is compressive failure, that is, the crushing of the building material. Since stone has tremendous compressive strength, however, it is not likely that the vaulting will fail in compression no matter how large the vaults. Therefore large fan vaults can be constructed on the same principles as smaller ones. Nevertheless, it took many decades for builders to understand that the stability of the fan vault is not compromised by its size. By the end of the 15th century this fact had been grasped and quite audacious ceiling designs were being attempted.

The construction of the largest fan vaults entailed some intriguing building techniques. The stone for the vaulting in the King's College chapel was taken by boat up the river Cam or by wagon to the building site. The rough stones from the quarries were in large blocks; a typical rough stone was probably a cube about two-thirds of a meter on a side. Temporary workshops were erected next to the building site and the stones were partially carved there.



CEILING TRACERY in the Chapel of Henry VII is among the most intricate examples of the patterns carved on the surface of fan vaults. The chapel was meant to glorify the Tudor dynasty; the carving is the material analogue of ornate praise for the Tudors. The vaulting is constructed entirely of jointed masonry, in which closely fitted blocks of stone compose the vaulting surface. Jointed masonry affords the best surface on which to carve elaborate designs.

Indeed, in the workshops the chapel vaulting was prefabricated. The rough stones were carved to the correct size and shape and the surfaces where the stones were to join were precisely finished. The rib and the tracery on the face of the block that was to be on the exposed surface of the conoid, however, needed only to be roughly carved. The stone was marked with a symbol indicating the quadrant of the bay it was to go into. Arabic or Roman numerals were carved at the end of the stone near the joint to designate the horizontal level of the conoid the stone was intended for, the exact location within that level and the timing of the insertion of the stone in the construction process.

The stone setter on the building site had only to observe the markings to put the stone in its correct place. The prefabrication of the vaulting stones may have required more skill than the actual construction. From a study of the construction markings visible on the upper surface of the conoids in the King's College chapel I have concluded that the bays in the main vault were built as follows. First the walls were erected and the roof was installed, following the common medieval building practice. Then the large transverse arches between bays were erected. All the stone setting probably required elaborate centering. The next step was the construction of the conoids, beginning at the long sides of the bays along the transverse arches and proceeding inward. One horizontal level at a time was constructed in each conoid. When the arches met in the center of the bay, the large stone that constitutes the keystone for both horizontal arches was placed. When the conoids were complete, the longitudinal ridge stones on each side of the central bosses were put in position. The final step was the insertion of the heavy central bosses.

The bosses were dropped in place from above. Because of their great weight they induced compression in the vaulting, which slightly changed its shape and allowed the centering to be readily taken down. When the centering had been removed, the final carving of the surface of the conoid was probably done. After any cracks that might have appeared were mortared the conoids were sometimes painted. The vaulting in the chapel of King's College was originally to have been painted and gilded.

Although the main vault at King's College has the longest span of any fan vaulting, the most splendid fan vaulting is that in the Chapel of Henry VII. The span of the chapel, 10.6 meters, is only slightly less than that of King's College and the overall design is awesome, combining a simplicity of three-dimensional form with a very complex

and regular surface pattern. Because of the high degree of surface articulation, which was considered fitting for a monarch, the vaulting was constructed entirely of jointed masonry.

The structural solution in the Chapel of Henry VII is ingenious. Large, mostly hidden transverse arches support pendants placed near the wall. Conoids are built upward and outward from the pendants. At the pendants the transverse arch rises through the vaulting conoid and continues along the upper surface of the vaulting, disappearing from view in the central area of the bay. Practically the entire structural load is transmitted to the transverse arch at the pendant. The arch carries the thrust to the walls and buttresses. Hence the ceiling combines shell structures (the conoids) and frame structures (the transverse arches).

Unlike the conoids of many fan vaults, where each horizontal cross section is only a segment of a circle, the central conoids in the Chapel of Henry VII are fully circular in horizontal section. This is symbolically appropriate because the circle, representing the revolution of the heavens and the disk of the sun, was an important symbol in Tudor political iconography. In 1500, when Henry's chapel was designed, the Tudors were a young and self-conscious dynasty; Henry had seized power only 15 years before on the basis of a slim claim to the crown. The magnificently carved ceiling of the chapel with its multiple circular forms was intended as elaborate praise for a monarch anxious to secure a place for himself in the cosmic and historical orders. The ceiling tells us more about the chapel's patron and his society than it does about the personality of the designer, whose identity we can only guess at.

The ceiling of the Chapel of Henry VII represents the culmination of fan vaulting. In the 1540's Henry's son Henry VIII drastically reduced the authority and wealth of the monastic orders in whose buildings much fan vaulting had been included. The construction of fan vaulting came to a virtual halt for almost 100 years, until it was revived by the Church of England at the University of Oxford, which was then a stronghold of Anglicanism.

The construction of fan vaults has never ceased entirely. Some modern cathedrals have vaulting in imitation of the earlier styles. The later examples, however, lack the charm and interest of fan vaults built between 1350 and 1540. This interest comes not only from the beauty of the ceilings themselves, which is considerable, but also from the way the form was worked out by the builders, who were guided solely by aesthetic impulses and empirical results but who were nonetheless dramatically successful engineers.

Now NRI takes you inside the new TRS-80 Model III microcomputer to train you at home as the new breed of computer specialist!

NRI teams up with Radio Shack advanced technology to teach you how to use, program and service state-of-the-art microcomputers...

It's no longer enough to be just a programmer or a technician. With microcomputers moving into the fabric of our lives (over 250,000 of the TRS-80™ alone have been sold), interdisciplinary skills are demanded. And NRI can prepare you with the first course of its kind, covering the complete world of the microcomputer.

Learn At Home In Your Spare Time

With NRI training, the programmer gains practical knowledge of hardware, enabling him to design simpler, more effective programs. And, with advanced programming skills, the technician can test and debug systems quickly and easily.

Only NRI gives you both kinds of training with the convenience of home study. No classroom pressures, no night school, no gasoline wasted. You learn at your convenience, at your own pace. Yet you're always backed by the NRI staff and your instructor, answering questions and giving you guidance.

You Get Your Own Computer to Learn On and Keep

NRI training is hands-on training with practical experiments and demonstrations. You don't just program your computer, you go inside it... watch how circuits interact... interface with other systems ...gain a real insight into its nature.

You also work with an advanced liquid crystal display hand-held multimeter and the NRI Discovery Lab,® performing over 60 separate experiments. Both microcomputer and equipment come as part of your training for you to use and keep.

Computer Assisted Instruction

Your TRS-80 even helps train you. You receive 8 special lesson tapes in BASIC computer language. Using them in your microcomputer, you "talk" to it as you progress. Errors are explained, graphics and animation drive home key points. Within a matter of minutes, you'll be able to write simple programs yourself.

Send for Free Catalog... No Salesman Will Call

Get all the details on this exciting course in NRI's free, 100-page catalog. It shows all equipment, lesson outlines, and facts on other electronics courses such as Electronic Design, Industrial Electronics, TV/Audio/Video Servicing...11 different career opportunities in all. Send today, no salesman will ever bother you. Keep up with the latest technology as you learn on the latest model of the world's most popular computer. If coupon has been used, write to NRI Schools, 3939 Wisconsin Ave., Washington, D.C. 20016.

(TRS-80 is a trademark of the Radio Shack division of Tandy Corp.)

NRI McGraw-Hill

NRI Schools
McGraw-Hill Continuing Education Center
3939 Wisconsin Avenue
Washington, D.C. 20016
We'll give you tomorrow.

NO SALESMAN WILL CALL.

Name (Please Print) Age

Street

City/State/Zip

Accredited by the Accrediting Commission of the National Home Study Council

Please check for one free catalog only.
- ☐ Computer Electronics including Microcomputers
- ☐ Color TV, Audio, and Video System Servicing
- ☐ Electronics Design Technology
- ☐ Digital Electronics
- ☐ Communications Electronics • FCC Licenses • Mobile CB • Aircraft • Marine
- ☐ Industrial Electronics
- ☐ Basic Electronics
- ☐ Small Engine Servicing
- ☐ Appliance Servicing
- ☐ Automotive Servicing
- ☐ Auto Air Conditioning
- ☐ Air Conditioning, Heating, Refrigeration, & Solar Technology
- ☐ Building Construction

All career courses approved under GI bill.
☐ Check for details

183-023

THE AMATEUR SCIENTIST

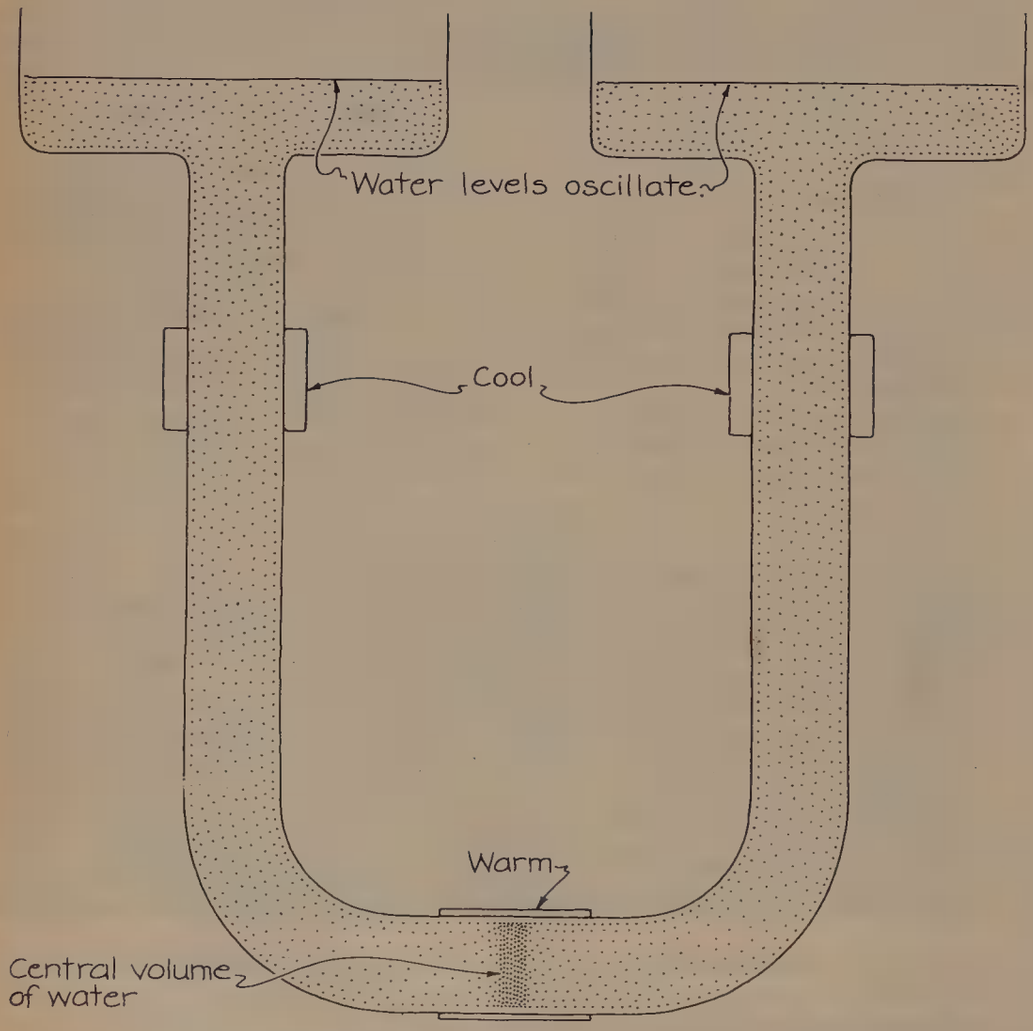
Thermal oscillators: systems that seesaw, buzz or howl under the influence of heat

by Jearl Walker

A thermal oscillator is a curiosity that turns up in physics classes because it is an example of a self-sustained, nonlinear oscillating system. What happens is that one part of the system vibrates continuously because of the transfer of heat. Here I shall discuss four such devices, two of them new and two that were demonstrated in the 19th century. Although each one depends on the periodic redistribution of heat, they are all quite different.

Pierre Welander of the University of Washington has devised a novel thermal oscillator based on water. The opening on each side of a U-tube is fitted with a wide container. The tube and about half of each container are filled with water. The tops of the containers are open. The bottom of the tube is heated; the upper sides are cooled. Soon the water level in the apparatus begins to oscillate between the two sides with a period of between several minutes and several hours.

Suppose the system is initially in complete equilibrium with its two sides containing equal amounts of water at the same temperature distribution. Consider a tiny volume of water lying within a thin slice through a cross section at the bottom of the tube. The volume is pressed from both sides because of the weight of the water in the two sides of the tube.



The physical conditions of a thermal oscillator devised by Pierre Welander

Two factors determine the weight of a column of water: the height of the column and the density of the water. The density is determined by the temperature: cooler water is denser than warmer water. Since initially the two sides of the system are identical in water level and temperature distribution, the central volume of water is pressed equally from both sides. It remains stationary and all the water is in equilibrium.

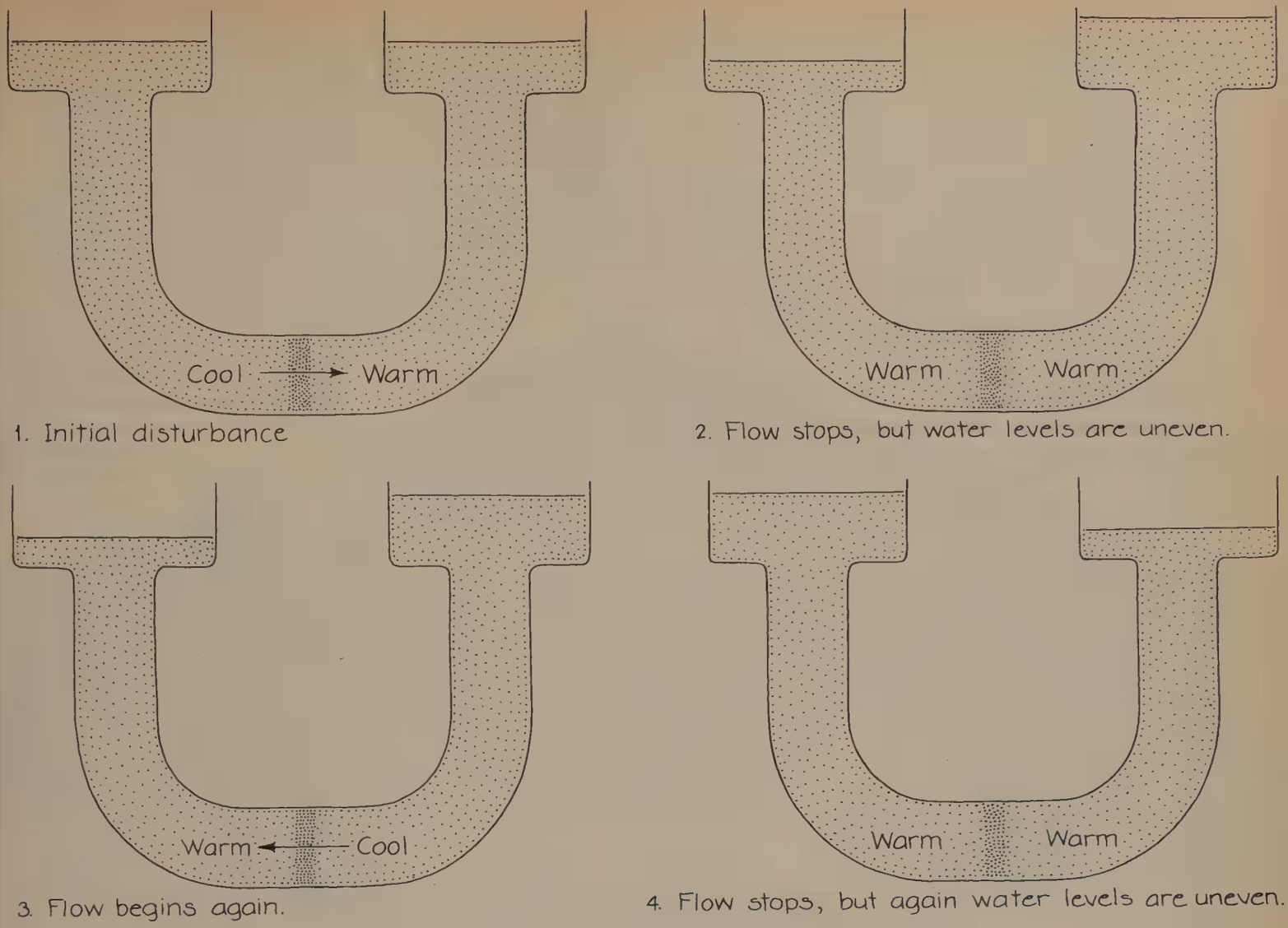
Suppose a disturbance from the environment causes a small amount of water to move from the left side to the right. The flow upsets the symmetry of the system; now the two sides differ in both water level and temperature distribution. The motion lowers the level on the left side and moves cooler water downward. It raises the level on the right side and moves warmer water to the right at the bottom of the U-tube.

If the containers on top of the U-tube are wide, the changes in the water levels are initially small. The broken symmetry of the temperature distribution is more important. Since the cooler water at the bottom left is denser than the warmer water at the bottom right, the central volume of water is no longer under equal pressure from the two sides. The left side now pushes more. The uneven pressures tend to push more water to the right.

Although the initial change in water levels is small, the continued flow increases the difference until it is significant. Eventually the right side applies more pressure on the central volume than the left side because of the difference in the water levels. The flow from left to right is slowed and then stopped. The process is gradual, however, and so the heating and cooling of the tube has a chance to reestablish the temperature symmetry. By the time the flow stops the system is again approximately symmetrical in temperature, but now the level is higher on the right side than on the left. The water begins to flow back to the left side, and the cycle is renewed in the opposite direction. The system remains in oscillation.

The oscillating flow continues indefinitely because of the heating and cooling of the tube. If the tube is allowed to reach a uniform temperature, the oscillation soon damps down into a stable situation with equal water levels on the two sides. As long as the heating and cooling are continued, however, the oscillation is maintained. Every time the system passes through the state where the two water levels are equal the imbalance in densities drives the water until the levels are unequal, continuing the oscillation.

The dimensions of the system are critical to the onset of oscillation. The top

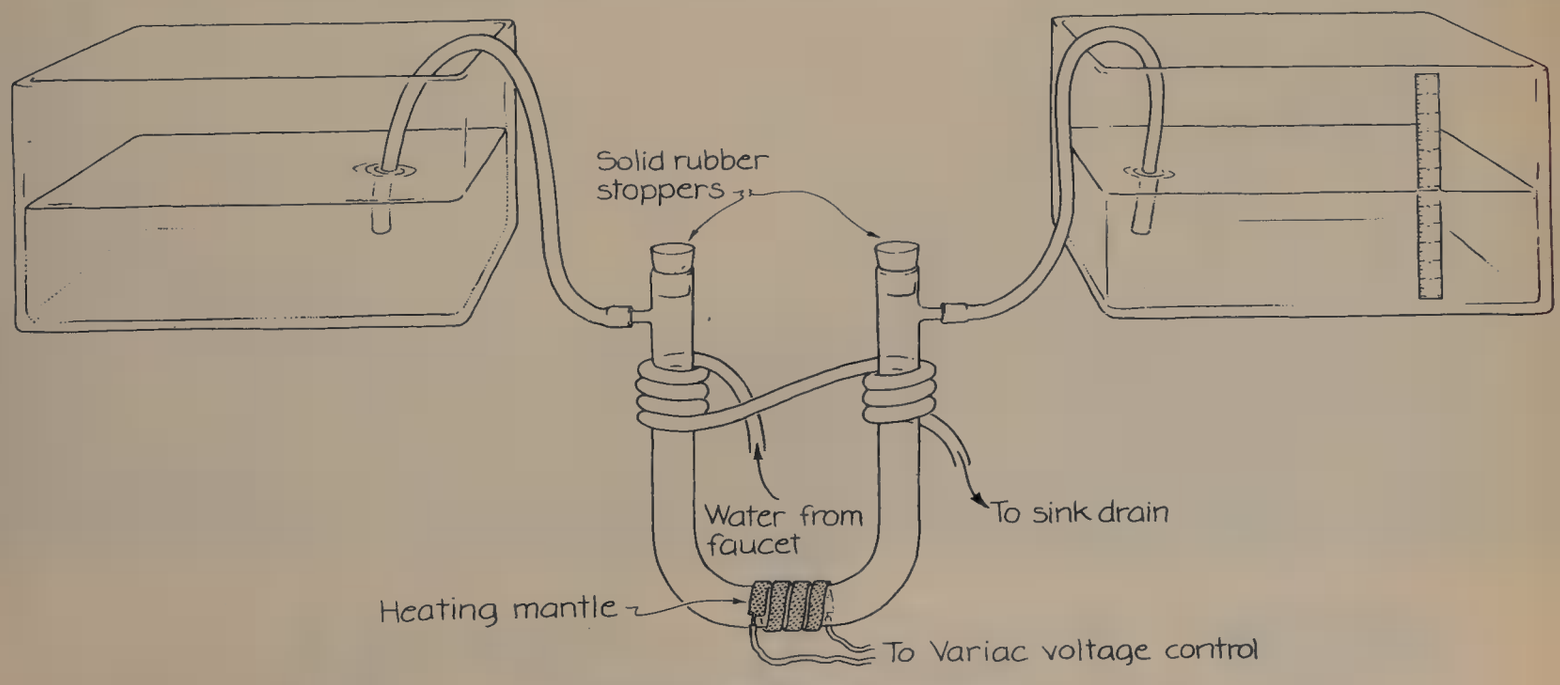


The pattern of flow in a fluid oscillator

surface area of each container must be larger than the cross-sectional area of the tube by an amount that is related to the temperature distribution in the water. The ratio of the area of a container to the area of the tube must exceed 5,000 divided by the maximum temperature difference (in degrees Celsius) that is set up in the water by the heating and cooling.

For example, if the system gives rise to a temperature difference of 50 degrees C., the top surface area of the container should be at least 100 times larger than the cross-sectional area of the tube. A smaller temperature difference demands an even larger ratio of the areas. If the containers are unusually large, of course, the oscillation of the water levels may be difficult to perceive.

The requirement for the relative sizes



A setup for an oscillator similar to Welander's

of the areas arises from the interplay of pressure on the central volume at the bottom of the U-tube. Suppose the initial disturbance from the environment moves a small amount of water to the right. The resulting asymmetry in temperature tends to drive more water to the right. This flow is countered by the difference in water levels on the two sides. If the top surface area of the containers is too small, the flow immediately creates a difference in water levels that is more than enough to stop the flow resulting from the asymmetry in temperature. Equilibrium is quickly reestablished.

If oscillation is to begin, the top surface area of the containers must be relatively large. That way the flow driven by the temperature asymmetry does not immediately yield a large difference in water levels. The flow has time to develop. The period of oscillation is set by the relative areas of the containers and the tube, the length of the tube, the viscosity of the water and the acceleration of gravity.

An apparatus similar to Welander's can be made by fusing glass containers on top of a U-tube. Since the tube should be Pyrex to preclude its shattering from being heated and cooled, fusing containers onto the tube seemed too difficult to me. I was also worried about how to mount such a delicate system so that it would stay upright. I chose a U-tube with side arms near the top. Short lengths of rubber hose connected the arms to plastic containers I took from my kitchen. The containers were identical; the top surface area of each one was about 200 times the cross-sectional area of the U-tube.

I taped a transparent ruler to the inside of one container so that I could monitor the water level. The U-tube and about half of each container were filled with water. Around the bottom of the U-tube I wrapped a heating mantle of the type that can be found among the supplies of a standard chemical laboratory. This mantle contains a high-resistance electrical wire that supplies heat. I plugged it into a Variac voltage control so that I could regulate the amount of heat applied to the tube. (I took great pains to make sure there would be no sudden leak that would let water reach the electrical line.)

For cooling I wrapped a thin-walled rubber hose several times around the top of the U-tube. Tap water could then be run through the hose. Although the hose did help to cool the water in the top of the U-tube, I decided a better system was needed. The conduction of heat through the glass wall of the U-tube and the rubber wall of the hose seemed to be too slow.

I had no way of measuring the temperature distribution of the water in the U-tube. After experimenting with several settings on the Variac I did manage to observe small oscillations in the water level of the containers. The level adjacent to the ruler rose and fell by about a millimeter every five minutes. Although the oscillations were not large or always regularly spaced, they were of the type observed by Welander because they disappeared when the heating and cooling systems were eliminated.

Welander told me about a more sensitive monitor of the water level. A small cork is hung from the edge of the container on a short arm made of wire. The cork floats on the water. A vertical straw is mounted on the cork. When the water level changes, the floating cork rotates about the wire arm. The upper end of the straw magnifies the motion.

Welander has also described a different kind of thermal oscillator in which a thin layer of ice forms periodically in a container of water. His experiment was done with about 50 liters of water in a Plexiglas container having a square bottom 50 centimeters on an edge. The container was thermally insulated. Near the bottom a heating element released about 100 watts of heat. Magnetic stirrers constantly mixed the water to distribute the heat. The apparatus was kept in a compartment where the air temperature was maintained at −20 degrees C. The air above the water was also continuously stirred.

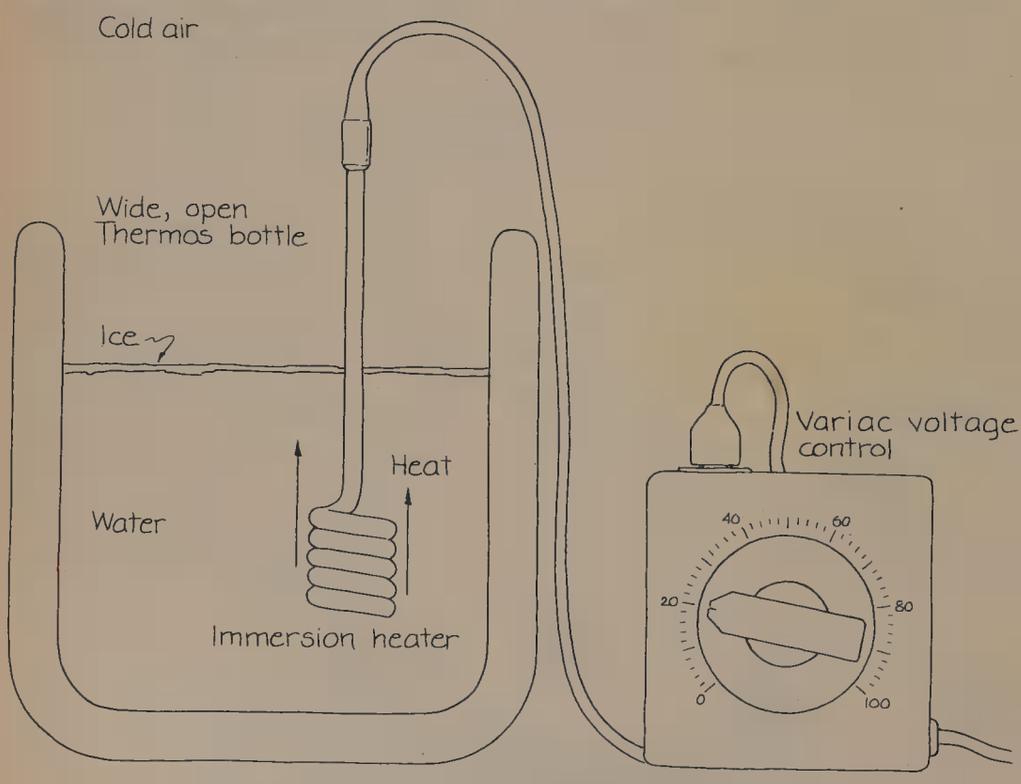
Welander monitored the temperature of the water at half of the depth. He also watched for the formation of ice. The surface periodically froze and melted as the temperature of the water varied by about 10 degrees. The water temperature rose to about 10 degrees C. during the phase of the oscillation when the top surface was frozen. It fell to almost the freezing point during the phase without ice. The period of oscillation was roughly 50 hours.

You can observe similar oscillations of freezing and melting water at home. Welander suggests using a Thermos bottle with a wide top. Fill it with water almost to the top, insert a heating coil below the surface and place the apparatus in a freezer. Ice periodically forms and disappears, although the oscillations are not as uniform as those Welander observed in the larger apparatus.

I followed his procedure with one change. The heating coil was the immersion heater with which I normally heat water for instant coffee. The heater was suspended in place over the center of the Thermos by several strips of reinforced tape. A problem was that the metal casing of the heater extended upward through the surface of the water. Heat from it surely interfered with the periodic formation of ice.

I plugged the immersion heater into a Variac so that I could experiment with the rate at which heat was released in the water. The rubber gasket lining the edge of the freezer door compressed enough so that I could close the door over the cord. If it had not closed completely, I was prepared to pack the edge with cloth to keep out room air. I turned the control of the freezer to its coldest setting. Again I took precautions to keep water from spilling onto the electrical supply.

The oscillations in the formation of ice are due to the poor transfer of heat through ice compared with the transfer through liquid water that is being stirred, either mechanically or by convection. The heat supplied to the system



An arrangement in which the oscillation is between ice and liquid water

What is needed in a time when people are keeping their cars longer, is a car people can be proud of longer. That's why, in 65 plants across the country, over 1000 Employee Involvement Groups have been established by Ford Motor Company and the UAW.

Their goal? Build the highest quality cars and trucks in the world. So you'll buy the results of their labor instead of someone else's. For example, members of this group at the Ford Lorain Assembly Plant set out to protect the 1983 Cougar's aerodynamic new body. Months before the first car was due to roll off the assembly line. They knew if their new Cougar was going to survive the rigors of the real world, it had to have a tough skin. So working with Ford management and engineers,

Lorain employees used their on-the-job experience to help set up new computer-guided spray equipment. A system specifically designed to apply paint and primer more uniformly to the Cougar's dramatic new shape. Thus enhancing the quality of the car's finish as well as its looks. And that's only one example of 129 employee-inspired quality improvements adopted by Ford management on the new 1983 Cougar. Employees speaking and management listening. Another reason Ford Motor Company achieved the highest quality rating of any major American car or truck maker. A fact based on a national survey measuring things gone wrong conducted among 11,000 people owning Ford and competitive 1982 new vehicles.

Flaw Fighters.

Get it together—buckle up

Ford
Mercury
Lincoln
Ford Trucks
Ford Tractors

Quality is Job 1.

Ford

Over 300 manufacturers sell their microcomputers with Microsoft software. There's a reason.

Powerful simplicity. That's the concept behind every software program we write. Powerful programs that let you spend more time thinking about the problem...and less time thinking about the computer. That concept has earned us the confidence of over 300 microcomputer manufacturers. Starting with the very first microcomputer you could buy. And today, Microsoft® languages, operating systems and applications software are running on well over a million microcomputers. Worldwide.

Made for each other. Manufacturers literally build their computers around Microsoft software. That's no exaggeration. In fact, Microsoft frequently participates in the initial design and development of microcomputers. That's a major reason why Microsoft software runs so well on the majority of the world's 8- and 16-bit systems.

User-oriented. Microsoft is the only software supplier to offer a full range of compatible operating systems, languages, utilities and applications programs. If you're not a computer expert, here's what that means to you: Better programs. Programs that are not only more powerful, but easier to learn and use.

Better tools. We can't buy your trust. We have to earn it. With better products. Tools you can easily use to solve complex problems. We started with the first BASIC for the first microcomputer. Today, we offer a broad range of proven tools for microcomputers. Including SoftCard™ and IBM® RAMCard,™ products that materially increase the capabilities of Apple ® and IBM PC computers. Plus, business and management software such as Multiplan,™ the powerful electronic worksheet you can learn to use in just a few hours. Better tools. Because they're designed specifically for your computer.

Ask your Microsoft dealer. Most microcomputer manufacturers offer their systems with some Microsoft software, but you'll undoubtedly want more. Programs and languages that solve your specific problems. Programs for your industry, business or home. Programs that do more, yet ask less. Ask your computer software dealer for a demonstration. You'll see why more than 300 microcomputer manufacturers offer their systems with Microsoft software. The reason is powerful simplicity.

BETTER TOOLS FOR MICROCOMPUTERS

MICROSOFT™

MICROSOFT CORPORATION
10700 NORTHUP WAY
BELLEVUE, WASHINGTON 98004

Microsoft is a registered trademark, and Multiplan, SoftCard, RAMCard and the Microsoft logo are trademarks of Microsoft Corporation. Apple and the Apple logo are registered trademarks of Apple Computer, Inc. IBM and the IBM logo are registered trademarks of International Business Machines Corporation. The Radio Shack logo is a registered trademark of Radio Shack, a Division of Tandy Corporation. The Burroughs logo is a registered trademark of Burroughs Corporation. The Victor logo is a registered trademark of Victor Business Products. The Wang logo is a registered trademark of Wang Laboratories, Inc.

by the submerged coil must travel upward because the sides of the container are insulated. The surface of the water is being cooled by air at a temperature below the freezing point of water.

The system can develop into either of two possible steady states. In one of them the surface of the water remains liquid because the air is not particularly cold and the supply of heat from the coil always prevents freezing. The heat travels upward by conduction and convection in the water. The transfer to the top surface is sufficiently fast to prevent freezing.

In the other steady state the system permanently freezes at the top because the air is quite cold and the rate at which heat is supplied is not particularly high. The heat from the coil travels upward to the ice layer by conduction and convection. There it is conducted slowly through the ice and carried away by the convection currents in the air. The system is in a steady state because although the conduction through the ice is slow, it is still fast enough to keep the temperature of the underlying water from rising high enough to melt the ice.

Oscillations can begin if the air is quite cold and the rate at which heat is supplied by the coil is high. Suppose the water is not frozen. The cold air removes heat from the surface fast enough to initiate freezing. When the ice layer forms, the removal of heat from the system diminishes because ice does not transfer heat as well as stirred liquid water. Thus the temperature of the water below the ice begins to rise. Eventually the water is warm enough to melt the ice.

When the ice disappears, the rate of heat transfer increases and the temperature of the water drops. Later the surface temperature is again at the freezing point and a fresh layer of ice forms.

A different kind of thermal oscillator was discovered in 1805 by an inspector in a smelting factory in Saxony. When a heated bar of one metal is laid across a block of another metal, the bar may oscillate vigorously and noisily. This oscillator was rediscovered independently about 25 years later by Arthur Trevelyan and has been known since as the Trevelyan rocker.

I have seen two kinds of demonstration of this type of thermal oscillator. In each one the choice of metals is important. In the first demonstration two lead plates are mounted in a vise. The top edges of the plates projecting above the vise are filed sharp so that there is a slight separation between them. Then a brass bar is balanced across the plates. When the bar is heated, it begins to rock back and forth. As it hits the edge of one plate and then the edge of the other there is an almost steady tapping.

In the second demonstration a piece of brass with a triangular cross section has two sharp edges at one angle of the triangle. The two edges rest at right angles on the edge of a block of lead. Extending from the piece of brass on a line with the two sharp edges is a brass rod that has a weight at the other end; near the far end the rod rests freely on a support. When the rod is heated, the sharp edges of the brass piece rock back and forth around the axis of the rod, making a sound that is approximately constant in frequency as they tap on the lead block.

The rod in Trevelyan's device was copper. Brass is more commonly used today. Perhaps certain other metals could serve in place of brass, copper or lead, but the choice must be made with care if the continuous oscillations are to be produced. The oscillations and the associated tone were apparently first explained by Sir John Leslie soon after the rocker was demonstrated by Trevelyan. Recent work has refined the explanation but has not replaced it.

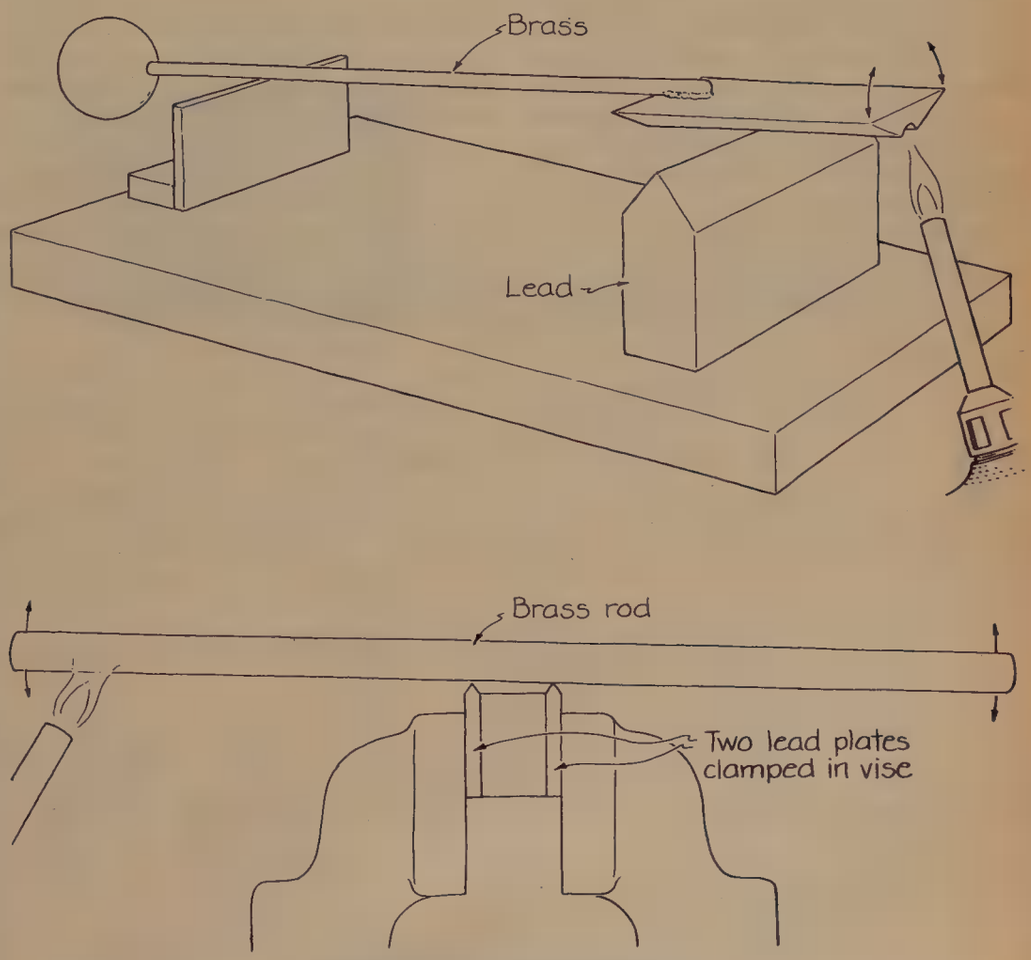
Consider the demonstration in which a brass piece rests on a lead block. The regions of contact are the two sharp edges of the brass piece. As the brass rod and the brass piece are heated by a flame, heat is transferred to the lead through the two edges. The rate of transfer through an edge depends partly on the weight resting on it: more weight provides better contact and a greater transfer of heat. One of the edges initially transfers heat better. Thermal expansion builds up the underlying lead into a tiny mound. The expansion can be fast enough to push the edge upward and off the mound.

Two things happen. Since one edge has been pushed up, the weight of the brass piece is shifted to the other edge. The heat is now transferred to the lead at the second edge. The mound at the first edge cools and shrinks. A fresh mound rises at the second edge. The system oscillates as the cycle of rising and falling mounds repeats.

The rate of tapping depends partly on the dimensions of the rocker. With some rockers the rate is so high that a continuous tone is heard.

Why do some materials not work? One reason is that the block must be of a material that expands when it is heated. Lead expands well, even more than brass and copper do. A second reason is that the vibrations of the rocker can be maintained only if the heat transferred to the block under an edge is not rapidly dispersed. If it is conducted away too fast, the region of the block just under the edge will not have enough time for thermal expansion. Lead conducts just poorly enough to ensure that it is effective for the purpose.

A third reason is that the block must conduct heat fast enough to keep the heat from remaining too long in the region under an edge, otherwise the thermal expansion would persist too long and the mound would subside too slow-



Two versions of the Trevelyan rocker

ly. The edge would no longer be lifted off the block.

You might like to reverse the brass and lead so that a lead piece with two sharp edges rests on a block of brass. The heating procedure cannot be reversed; you must still heat the brass piece. You might also enjoy monitoring the oscillations in either system by flashing a stroboscopic light on the rocker. Although the motion is slight, it can be amplified by attaching a needle or a thin rod to the rocker at right angles to its long axis. The end of the needle moves through a greater distance than the rocker and so makes observation of the motion easier.

Another thermal oscillator was once a standard demonstration in physics classes. A section of wire mesh is pushed about a quarter of the way up inside a vertical pipe. The mesh is bent at the edges so that it remains in place inside the pipe. A burner is set under the pipe so that both the pipe and the mesh get hot. When the burner is removed, the pipe howls for several seconds. The tone is nearly pure because it is the fundamental frequency for acoustic oscillations inside the pipe.

This system was discovered in 1859 by Pieter Leonhard Rijke of the University of Leiden. He thought the sound resulted from thermal contraction of the wire mesh and pipe as they cooled. He believed oscillations accompanying the contractions excite the acoustic wave emitted by the pipe. Rijke was wrong; neither the mesh nor the pipe oscillates.

The correct explanation was supplied by Lord Rayleigh in 1878. The sound from the pipe arises from the forced oscillation of the air inside the pipe. When the burner is in place, it heats the air in the pipe and forces a convection current upward through the mesh and pipe. No sound is emitted at this stage because the air is flowing steadily upward.

When the heat source is removed, the convection current continues for a while because both the pipe and the mesh are hot. The many disturbances of the air in the pipe tend to create an acoustic standing wave in that air. I shall first explain such a standing wave in an unheated pipe and then show how the convection current in the Rijke pipe strengthens it until the pipe howls.

An acoustic standing wave inside a straight pipe can be created by the interference of two sound waves traveling in opposite directions through the pipe. The sound waves must be identical except for their direction of travel. They pass through the pipe, reflect from the open ends and then pass through the pipe again. As they pass through each other their interference forces the molecules of air to oscillate by a small amount parallel to the length of the pipe. Because of the repeated interference of the waves a pattern develops in the oscillation of the molecules. The molecules at some places along the length of the pipe never move. Such a place is known as a displacement node. Molecules at other places oscillate by some maximum amount. Such a place is known as a displacement antinode.

This pattern of repeated interference of two traveling waves is called a standing wave because the positions of the nodes and antinodes in the pipe are stationary. The simplest standing wave in the pipe is termed the fundamental or the first harmonic. Its pattern consists of a displacement node halfway through the pipe and an antinode at each end. Hence the molecules near the open ends vibrate parallel to the length of the pipe by a maximum amount, whereas the molecules near the halfway point vibrate little or not at all.

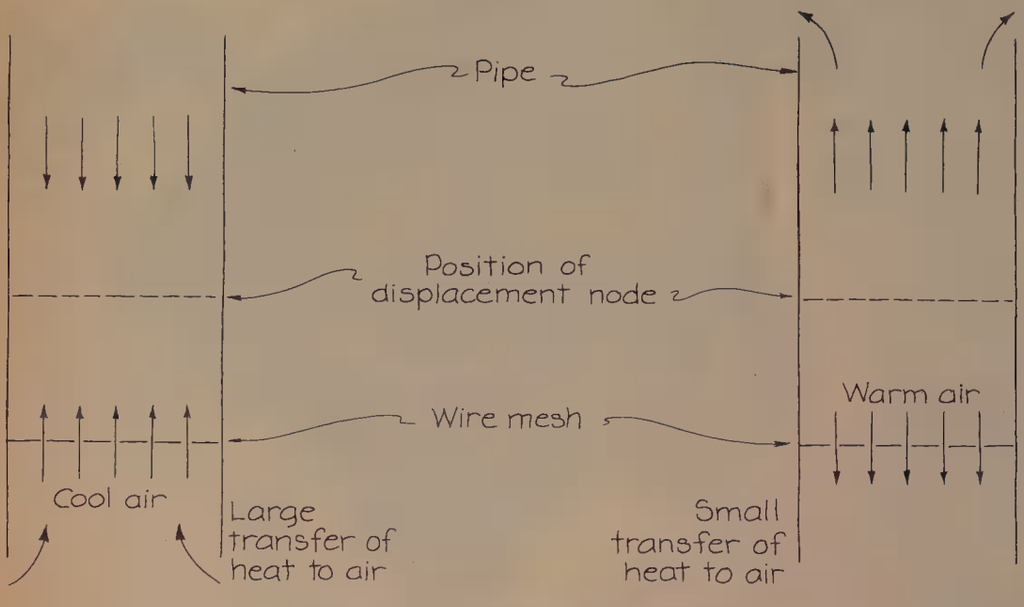
To excite the fundamental in a pipe the traveling waves of sound must have a frequency equal to the speed of sound divided by twice the length of the pipe. Once a standing wave has been initiated it can grow in strength if more energy is given to the sound waves. The pipe is then said to resonate at the frequency of the fundamental standing wave. The sound emerging from the pipe can be quite loud when resonance is achieved.

An acoustic standing wave inside a pipe also sets up periodic variations in the air pressure along the length of the pipe. Consider the node at the midpoint of the pipe when the fundamental is generated. On each side of the node molecules oscillate parallel to the length of the pipe. During one part of the oscillation the molecules on the opposite sides of the node move toward the node, increasing the air pressure there. Half a cycle later in the oscillations the molecules move away from the node, leaving low pressure there. As the molecules oscillate, the pressure at the node varies above and below the atmospheric pressure outside the pipe. The greatest variation in the pressure inside the pipe is at the displacement node.

In some physics textbooks the nodes and antinodes of an acoustic standing wave refer to the pressure variations. Thus a pressure antinode (a large variation in pressure) is found at a displacement node and a pressure node (no variation in pressure) is found at a displacement antinode. In Rayleigh's explanation of the Rijke tube the key is the relative positions of the wire mesh and the displacement node (the pressure antinode) at the midpoint of the pipe. The pipe must be vertical so that the heat forces a convection through it. The mesh must be in the lower half of the pipe, not too close either to the lower end or to the displacement node at the midpoint.

Assume the various disturbances of the air inside the pipe as it cools somehow initiate weak traveling waves that generate the pipe's fundamental standing wave. Air oscillates through the wire mesh. During part of the oscillation cycle the air must move upward through the mesh and toward the displacement node, building up the pressure at the node. During the other part of the cycle the air must move downward through the mesh and away from the node, leaving low pressure at the node.

Rayleigh argued that the oscillation of the air inside the pipe is aided by a transfer of heat from the hot mesh and pipe to the air. The transfer is efficient and beneficial, however, during only part of the oscillation cycle. When the air moves upward through the mesh toward the node, relatively cool air flows into the pipe through the bottom end. Because the temperature difference between the air and the mesh is then large, the mesh transfers a lot of heat to the air. That air is more buoyant because of its increased temperature. The heat supplied by the mesh effectively pushes the mass of air upward toward



How air moves through the Rijke-pipe oscillator

the node, increasing the pressure build-up there and furthering the oscillation.

During the downward motion of the air in the oscillation cycle the heat transferred to the air tends to buoy it upward. This time the push is in the wrong direction to aid the oscillation. Because the push is small, however, it does not seriously hinder the acoustic vibration. The air moving downward is almost as hot as the mesh because it has already been through that region. Hence the transfer of heat from the mesh to the air is small. What transfer does occur can only diminish the oscillation because an addition of heat to the air tends to drive it upward, not downward in synchrony with the oscillation.

The oscillation of the air in the pipe's fundamental standing wave is strengthened by the asymmetry in the transfer of heat during the two parts of the oscillation cycle. The oscillation receives a large push during one part and essentially no push during the other. Rayleigh likened this asymmetry to a force applied periodically to a swinging pendulum. Suppose you push on a pendulum every time it passes through its lowest point. In one part of the swing the push adds energy because you are pushing in the direction of motion. On the next pass of the pendulum through its lowest point, however, you are pushing in the direction opposite to the motion. You remove energy. As long as you apply the same force in each case the pendulum receives no net supply of energy. Its swing eventually decreases because of frictional losses of energy. To build up the swing you must push only when you can add energy to the oscillation.

There is another advantage in transferring heat to the air while the air moves toward the midpoint of the pipe. During the increase in pressure of the air near the midpoint the temperature and pressure of the air are proportional. When the hot mesh increases the temperature, it also enhances the pressure increase at the midpoint. When the air begins to move away from the midpoint and so decreases the pressure there, any addition of heat to the air would again tend to increase the pressure. Hence the fact that the transfer is small during this stage is advantageous to the oscillation.

You can demonstrate the howling of a Rijke pipe with almost any pipe that will not melt when it is heated. Window screen serves well for a mesh. A double layer may work better than a single one. Heat can be supplied by a laboratory burner, a kitchen stove or even a candle.

Once the howling begins try to move the pipe to a horizontal position and then restore it to the vertical. The natural convection through the pipe is destroyed when the pipe is horizontal; the sound ceases immediately. If the pipe is put vertical again before the mesh has cooled, the howling begins anew.

To preserve your copies of

SCIENTIFIC AMERICAN

A choice of handsome and durable library files—or binders—for your copies of SCIENTIFIC AMERICAN. Both styles bound in dark green library fabric stamped in gold leaf.

Files **Each file holds 12 issues. Price per file $5.95; three for $17.00; six for $30.00, postpaid.**
(Add $2.50 each outside U.S.A.)

Binders **Each binder holds 12 issues. Issues open flat. Price per binder $7.50; three for $21.75; six for $42.00, postpaid.**
(Add $2.50 each outside U.S.A.)

To: *Jesse Jones Box Corp., P.O. Box 5120, Philadelphia, Pa. 19141*

file binder

Send me_____SCIENTIFIC AMERICAN
☐ Files ☐ Binders
For issues dated ☐ through 1982 ☐ 1983 or later.
I enclose my check or money order for $_____ *(U.S. funds only).*

Name _____
(please print)

Address _____

City _____

State _____ **Zip** _____

NOTE: Satisfaction guaranteed or money refunded. Allow four to six weeks for delivery.

BIBLIOGRAPHY

Readers interested in further explanation of the subjects covered by the articles in this issue may find the following lists of publications helpful.

METAMAGICAL THEMAS

LISP 1.5 PROGRAMMER'S MANUAL. John McCarthy. The MIT Press, 1962.

THE LITTLE LISPER. Daniel P. Friedman. Science Research Associates, Inc., 1974.

LET'S TALK LISP. L. S. Siklóssy. Prentice-Hall, Inc., 1976.

LISP. Patrick Henry Winston and Berthold K. Horn. Addison-Wesley Publishing Co., Inc., 1981.

THE FUTURE OF AMERICAN AGRICULTURE

FARM AND FOOD POLICY: ISSUES OF THE 1980S. Don Paarlberg. University of Nebraska Press, 1980.

THE FUTURE OF AMERICAN AGRICULTURE AS A STRATEGIC RESOURCE. Edited by Sandra S. Batie and Robert G. Healy. The Conservation Foundation, 1980.

THE LATTICE THEORY OF QUARK CONFINEMENT

THE CONFINEMENT OF QUARKS. Yoichiro Nambu in *Scientific American,* Vol. 235, No. 5, pages 48–60; November, 1976.

THE BAG MODEL OF QUARK CONFINEMENT. Kenneth A. Johnson in *Scientific American,* Vol. 241, No. 1, pages 112–121; July, 1979.

GAUGE THEORIES OF THE FORCES BETWEEN ELEMENTARY PARTICLES. Gerard 't Hooft in *Scientific American,* Vol. 242, No. 6, pages 104–138; June, 1980.

LATTICE GAUGE THEORIES AND MONTE CARLO SIMULATIONS. Claudio Rebbi in *Non-Perturbative Aspects of Quantum Field Theory.* World Scientific Publishing Co., Singapore, 1982.

SYNTHETIC VACCINES

CHEMICAL SYNTHESIS OF A POLYPEPTIDE PREDICTED FROM NUCLEOTIDE SEQUENCE ALLOWS DETECTION OF A NEW RETROVIRAL GENE PRODUCT. J. G. Sutcliffe, T. M. Shinnick, N. Green, F.-T. Lie, H. L. Niman and R. A. Lerner in *Nature,* Vol. 287, No. 5785, pages 801–805; October 30, 1980.

ANTIBODIES TO CHEMICALLY SYNTHESIZED PEPTIDES PREDICTED FROM DNA SEQUENCES AS PROBES OF GENE EXPRESSION. Richard A. Lerner, J. Gregor Sutcliffe and Thomas M. Shinnick in *Cell,* Vol. 23, No. 2, pages 309–310; February, 1981.

TAPPING THE IMMUNOLOGICAL REPERTOIRE TO PRODUCE ANTIBODIES OF PREDETERMINED SPECIFICITY. Richard A. Lerner in *Nature,* Vol. 299, No. 5884, pages 592–596; October 14, 1982.

THE OPTICAL COMPUTER

OPTICAL BISTABLE DEVICES: THE BASIC COMPONENTS OF ALL-OPTICAL SYSTEMS? H. M. Gibbs, S. L. McCall and T. N. C. Venkatesan in *Optical Engineering,* Vol. 19, No. 4, pages 463–468; July/August, 1980.

OPTICAL BISTABILITY. Edited by Charles M. Bowden, M. Ciftan and H. R. Robl. Plenum Press, 1981.

DYNAMIC NON-LINEAR OPTICAL PROCESSES IN SEMICONDUCTORS. Alan Miller, David A. B. Miller and S. Desmond Smith in *Advances in Physics,* Vol. 30, No. 6, pages 697–800; November/December, 1981.

BISTABLE OPTICAL DEVICES: PHYSICS AND OPERATING CHARACTERISTICS. David A. B. Miller in *Laser Focus with Fiberoptic Technology,* Vol. 18, No. 4, pages 79–84; April, 1982.

HIDDEN VISUAL PROCESSES

EYE TORSION AND VISUAL TILT ARE MEDIATED BY DIFFERENT BINOCULAR PROCESSES. Jeremy M. Wolfe and Richard Held in *Vision Research,* Vol. 19, No. 8, pages 917–920; 1979.

CYCLOPEAN STIMULATION CAN INFLUENCE SENSATIONS OF SELF-MOTION IN NORMAL AND STEREOBLIND SUBJECTS. Jeremy M. Wolfe and Richard Held in *Perception and Psychophysics,* Vol. 28, No. 2, pages 139–142; 1980.

INTEROCULAR TRANSFER OF VISUAL AFTEREFFECTS. R. Blake, R. Overton and S. Lema-Stern in *Journal of Experimental Psychology: HPP,* Vol. 7, pages 367–381; 1981.

IS ACCOMMODATION COLORBLIND? FOCUSING CHROMATIC CONTOURS. Jeremy M. Wolfe and D. Alfred Owens in *Perception,* Vol. 10, pages 53–62; 1981.

A PURELY BINOCULAR MECHANISM IN HUMAN VISION. Jeremy M. Wolfe and Richard Held in *Vision Research,* Vol. 21, No. 12, pages 1755–1759; 1981.

THE ACTIVE SOLAR CORONA

THE SOLAR CORONA. Jay M. Pasachoff in *Scientific American,* Vol. 229, No. 4, pages 68–79; October, 1973.

A NEW SUN: THE SOLAR RESULTS FROM SKYLAB. John A. Eddy. Scientific and Technical Information Office, National Aeronautics and Space Administration, U.S. Government Printing Office, 1979.

DART-POISON FROGS

PRELIMINARY EVALUATION OF SKIN TOXINS AND VOCALIZATIONS IN TAXONOMIC AND EVOLUTIONARY STUDIES OF POISON-DART FROGS (DENDROBATIDAE). Charles W. Myers and John W. Daly in *Bulletin of the American Museum of Natural History,* Vol. 157, Article 3, pages 173–262; 1976.

CLASSIFICATION OF SKIN ALKALOIDS FROM NEOTROPICAL POISON-DART FROGS (DENDROBATIDAE). John W. Daly, George B. Brown, Monica Mensah-Dwumah and Charles W. Myers in *Toxicon,* Vol. 16, No. 2, pages 163–188; 1978.

MORPHOLOGY OF THE GRANULAR SECRETORY GLANDS IN SKIN OF POISON-DART FROGS (DENDROBATIDAE). Maria Neuwirth, John W. Daly, Charles W. Myers and Lois W. Tice in *Tissue & Cell,* Vol. 11, No. 4, pages 755–771; 1979.

TAXONOMY AND ECOLOGY OF *DENDROBATES BOMBETES,* A NEW ANDEAN POISON FROG WITH NEW SKIN TOXINS. Charles W. Myers and John W. Daly in *American Museum Novitates,* No. 2692, pages 1–23; January 29, 1980.

THE PHARMACOLOGY OF PUMILIOTOXIN-B, 1: INTERACTION WITH CALCIUM SITES IN THE SARCOPLASMIC RETICULUM OF SKELETAL MUSCLE. E. X. Albuquerque, J. E. Warnick, M. A. Maleque, F. C. Kauffman, R. Tamburini, Y. Nimit and J. W. Daly in *Molecular Pharmacology,* Vol. 19, No. 3, pages 411–424; May, 1981.

FAN VAULTING

MEDIEVAL STRUCTURE: THE GOTHIC VAULT. James J. Acland. University of Toronto Press, 1972.

FAN VAULTING: A STUDY OF FORM, TECHNOLOGY AND MEANING. Walter C. Leedy, Jr. Arts and Architecture Press, Santa Monica, 1980.

EXPERIMENTS IN GOTHIC STRUCTURE. Robert Mark. The MIT Press, 1982.

THE AMATEUR SCIENTIST

HEAT-MAINTAINED SOUNDS. E. G. Richardson in *Sound: A Physical Text-book.* Edward Arnold & Co., 1935.

THEORY OF RESONATORS. John William Strutt, Baron Rayleigh, in *The Theory of Sound.* Dover Publications, Inc., 1945.

OBSERVATION OF OSCILLATORY ICE STATES IN A SIMPLE CONVECTION EXPERIMENT. Pierre Welander in *Journal of Geophysical Research,* Vol. 82, No. 18, pages 2591–2592; June 20, 1977.

Marygrove College L
Detroit, Michigan 48
PLEASE DO NOT REM

SCIENTIFIC AMERICAN

MEMORY IN BIRDS

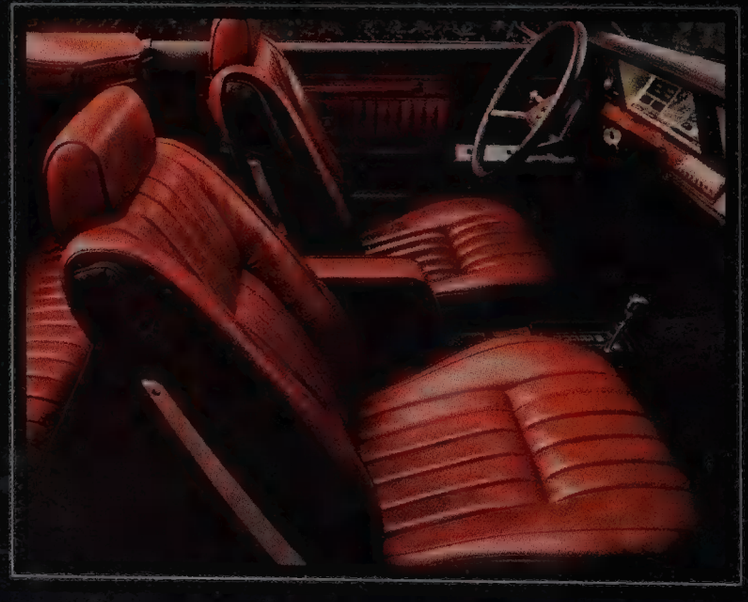
$2.50

March 1983

0 74820 08715 03

THE CHRYSLER LEBARON SERIES

For 1983, Chrysler re-introduces an American Classic. The Town & Country Convertible.



Established 1845 **SCIENTIFIC AMERICAN** March 1983 Volume 248 Number 3

ARTICLES

42 **THE LAW OF THE SEA, by Elisabeth Mann Borgese**
In December 119 nations signed a convention for a new international order of the ocean

50 **MICROPROGRAMMING, by David A. Patterson**
Most computers are controlled not by "hard-wired" circuitry but by a stored micropogram

58 **THE SCRIPT OF THE INDUS VALLEY CIVILIZATION, by Walter A. Fairservis, Jr.**
One of the four earliest civilizations has a script that is only now beginning to be deciphered

78 **MITOCHONDRIAL DNA, by Leslie A. Grivell**
The organelle has its own genetic system, different from all others, and even its own genetic cod

90 **THE FUTURE OF THE UNIVERSE, by Duane A. Dicus, John R. Letaw, Doris C. Teplitz and Vigdor L. Teplitz** A cosmological forecast of events through the year 10^{10}

102 **MEMORY IN FOOD-HOARDING BIRDS, by Sara J. Shettleworth**
One species remembers where it has put thousands of seeds for as long as several month

112 **OSCILLATING CHEMICAL REACTIONS, by Irving R. Epstein, Kenneth Kustin, Patrick De Kepper and Miklós Orbán** In them the concentrations of reactants rise and fal

124 **AUTOPSY, by Stephen A. Geller**
The procedure is profoundly informative, but it is now done in only 15 percent of the death

DEPARTMENTS

11 LETTERS
13 50 AND 100 YEARS AGO
20 THE AUTHORS
22 METAMAGICAL THEMAS
35 BOOKS
68 SCIENCE AND THE CITIZEN
138 THE AMATEUR SCIENTIST
146 BIBLIOGRAPHY

BOARD OF EDITORS Gerard Piel (Publisher), Dennis Flanagan (Editor), Brian P. Hayes (Associate Editor), Philip Morrison (Book Editor), Francis Bello, John M. Benditt, Peter G. Brown, Michael Feirtag, Diana Lutz, Jonathan B. Piel, John Purcell, James T. Rogers, Armand Schwab, Jr., Joseph Wisnovsky

ART DEPARTMENT Samuel L. Howard (Art Director), Steven R. Black (Assistant Art Director), Ilil Arbel, Edward Bell

PRODUCTION DEPARTMENT Richard Sasso (Production Manager), Carol Hansen and Leo J. Petruzzi (Assistants to the Production Manager), Carol Eisler (Senior Production Associate), Karen O'Connor (Assistant Production Manager), Carol Albert, Lori Mogol, Martin O. K. Paul, Julio E. Xavier

COPY DEPARTMENT Sally Porter Jenks (Copy Chief), Debra Q. Bennett, Mary Knight, Dorothy R. Patterson

GENERAL MANAGER George S. Conn

ADVERTISING DIRECTOR C. John Kirby

CIRCULATION MANAGER William H. Yokel

SECRETARY Arlene Wright

SCIENTIFIC AMERICAN (ISSN 0036-8733), PUBLISHED MONTHLY BY SCIENTIFIC AMERICAN, INC., 415 MADISON AVENUE, NEW YORK, N.Y. 10017. COPYRIGHT © 1983 BY SCIENTIFIC AMERICAN, INC. ALL RIGHTS RESERVED. PRINTED IN THE U.S.A. NO PART OF THIS ISSUE MAY BE REPRODUCED BY ANY MECHANICAL, PHOTOGRAPHIC OR ELECTRONIC PROCESS, OR IN THE FORM OF A PHONOGRAPHIC RECORDING, NOR MAY IT BE STORED IN A RETRIEVAL SYSTEM, TRANSMITTED OR OTHERWISE COPIED FOR PUBLIC OR PRIVATE USE WITHOUT WRITTEN PERMISSION OF THE PUBLISHER. SECOND-CLASS POSTAGE PAID AT NEW YORK, N.Y., AND AT ADDITIONAL MAILING OFFICES. AUTHORIZED AS SECOND-CLASS MAIL BY THE POST OFFICE DEPARTMENT, OTTAWA, CANADA, AND FOR PAYMENT OF POSTAGE IN CASH. SUBSCRIPTION RATE: $21 PER YEAR IN THE U.S. AND ITS POSSESSIONS; $27 PER YEAR IN ALL OTHER COUNTRIES. POSTMASTER: SEND ADDRESS CHANGES TO SCIENTIFIC AMERICAN, 415 MADISON AVENUE, NEW YORK, N.Y. 10017.

"WE'RE MANAGING DIVE

—Charles L. Brow

I began my career with AT&T climbing poles during a high school vacation. Now, that business I joined 43 years ago is facing its greatest challenge: the agreement with the Justice Department requiring us to divest the 22 Bell System telephone companies by early 1984.

This is clearly the most complex reorganization job ever undertaken by the management of any business. But it is especially involved because of the complexity of our business. We provide a service critical to business and commerce, to government and education, to national defense, to individuals in their daily lives. We are embedded in the very life of the nation.

We're adapting our business to what the public expects.

I'd be less than truthful if I didn't say we have mixed feelings about breaking up a 100-year-old institution which has served the nation very well. On the other hand, we cannot live outside our times. And clearly, the times have changed. The American people really don't want monopoly, no matter how well regulated. They want competition and the choices that brings.

Technology has changed as well. Not too long ago, data processing was one thing, communications another. That's no longer true. As a result, many companies here and abroad have the technical know-how as well as the marketing resources to deliver high-quality communications products, services, and systems. Thus, the Bell System could not expect to operate and do business as it always has.

Never underestimate the resourcefulness and dedication of telephone people to figure out how to reach a goal.

TITURE, NOT DEMOLITION."

Chairman of the Board, AT&T

The task of divesting is vastly complicated. But we have the human talent. We have the engineering and scientific resources. And we have our pride.

While our companies will be radically changed, the basic aims of AT&T and the telephone companies will not change: to serve the communications needs of the nation, and to assure America its place as a leader in high technology worldwide.

Restructured by a Consent Decree with the Justice Department, the new AT&T will be able to develop and apply its technologies to the fullest.

In considering the Consent Decree proposal, we thought long and hard about whether the new structure would imperil our ability to serve. And of course we never would have accepted any terms that automatically would have degraded the quality of service to the nation. The Decree gives us the opportunity to build even better communications services and systems because it removes the arbitrary constraints which have limited us in the applications of our technologies.

We're eager to bring the benefits of all our innovations to the consumer in the competitive Information Age marketplace.

Both AT&T and the local phone companies have a bright future.

We're at the heart of the fastest growing, most promising industry in the country. The new AT&T—with its long-distance network and its new subsidiary, American Bell, plus Western Electric and Bell Labs—is in league with the future.

The telephone companies, which already provide a local communications system reaching virtually every home and business, are enhancing the quality and capability of their local lines so they can handle total information services: voice, video, data, even the new cellular mobile phone services. We are pledged to divest these companies in sound financial shape. We will keep that pledge.

This industry is where the opportunities and the excitement are.

If we have a truly competitive communications marketplace, with regulation only where it is needed, I believe AT&T and the divested telephone companies have a significant and constructive role to play in revitalizing the American economy and in maintaining and enhancing US technological leadership in communications.

The Bell System as we now know it will be no more. We will divest. But we are not demolishing the promise of tomorrow. That promise is alive and well. Bell System people are ready for new directions.

Bell System

SCIENTIFIC AMERICAN

CORRESPONDENCE

Offprints of more than 1,000 selected articles from earlier issues of this magazine, listed in an annual catalogue, are available at 75 cents each. Correspondence, orders and requests for the catalogue should be addressed to W. H. Freeman and Company, 4419 West 1980 South Salt Lake City, UT 84121. Offprints adopted for classroom use may be ordered direct or through a college bookstore. Sets of 10 or more Offprints are collated by the publisher and are delivered as sets to bookstores.

Photocopying rights are hereby granted by Scientific American, Inc., to libraries and others registered with the Copyright Clearance Center (CCC) to photocopy articles in this issue of SCIENTIFIC AMERICAN for the flat fee of $2 per copy of each article or any part thereof. Such clearance does not extend to the photocopying of articles for promotion or other commercial purposes. Correspondence and payment should be addressed to Copyright Clearance Center, Inc., 21 Congress Street, Salem, Mass. 01970. Specify CCC Reference Number ISSN 0036–8733/83. $2.00 + 0.00.

Editorial correspondence should be addressed to The Editors, SCIENTIFIC AMERICAN, 415 Madison Avenue, New York, N.Y. 10017. Manuscripts are submitted at the authors' risk and will not be returned unless they are accompanied by postage.

Advertising correspondence should be addressed to C. John Kirby, Advertising Director, SCIENTIFIC AMERICAN, 415 Madison Avenue, New York, N.Y. 10017.

Subscription correspondence should be addressed to Subscription Manager, SCIENTIFIC AMERICAN, P.O. Box 5969, New York, N.Y. 10017. The date of the last issue on your subscription is shown in the upper right-hand corner of each month's mailing label. For change of address notify us at least four weeks in advance. Please send your old address (if convenient, on a mailing label of a recent issue) as well as the new one.

Name

New Address

Street

City

State and ZIP

Old Address

Street

City

State and ZIP

THE COVER

The painting on the cover shows an experiment testing the memory of a marsh tit, a small British bird related to the North American chickadees (see "Memory in Food-hoarding Birds," by Sara J. Shettleworth, page 102). The bird is fitted with a hemispherical patch that covers one eye; then it is given sunflower seeds, which it hoards by poking each one into the moss that fills a large tray. Later the bird is allowed to search for the seeds. If the patch has not been moved, the bird remembers where the seeds are hidden. If the patch has been moved, however, so that it covers the eye on the opposite side of the head (the eye that viewed the storage sites), the result is quite different: the bird cannot remember. In the marsh tit each eye sends information to only one side of the brain. Thus it appears that visual memories in the marsh tit cannot be shunted from one side of the brain to the other. The experiment was done by David F. Sherry, John R. Krebs and Richard J. Cowie of the University of Oxford.

THE ILLUSTRATIONS

Cover painting by Enid Kotschnig

Page	Source
26–29	Edward Bell
31	Ilil Arbel
43	UN Photo
44–45	Andrew Tomko
46	Walken Graphics
48	Fred N. Spiess and Peter E. Lonsdale, Scripps Institution of Oceanography
49	Andrew Tomko
51	Hewlett-Packard Company (*top*), Ilil Arbel (*bottom*)
52	Zilog, Inc.
53	Internal Revenue Service, U.S. Department of the Treasury (*top*); Ilil Arbel (*bottom*)
54–57	Ilil Arbel
59	James P. Blair, courtesy of the National Museum of Pakistan
60	James P. Blair, courtesy of the National Museum of India (*top*); American Museum of Natural History (*bottom*)
61–65	Andrew Christie
66	McGuire Gibson, Oriental Institute, University of Chicago (*top*); Andrew Christie (*bottom*)
79	Ernst F. J. Van Bruggen, State University at Groningen (*top*); Annika C. Arnberg and Ernst F. J. Van Bruggen, State University at Groningen (*bottom*)
80–84	Bunji Tagawa
85	Annika C. Arnberg and Ernst F. J. Van Bruggen, State University at Groningen (*top*); Bunji Tagawa (*bottom*)
86–87	Bunji Tagawa
88	J. H. Hoeimakers, University of Amsterdam
91	Ann Savage, Royal Observatory Edinburgh, Coonabarabran, Australia
92–100	Walken Graphics
103–104	Tom Prentiss
105	Stephen B. Vander Wall, Utah State University
106–110	Tom Prentiss
113	R. F. Bonifield
114–122	Allen Beechel
123	R. F. Bonifield
125–135	Carol Donner
136	Ilil Arbel
139–145	Michael Goodman

DATSUN 200-SX
IT'S GOT PERFORMANCE DOWN TO A SCIENCE.

The sleek Datsun 200-SX is a scientific paradox. Although it burns gas "faster" than conventionally powered cars, the 200-SX gets even better gas-mileage ratings this year than last. And last year's ratings were already remarkably high for a car that won two dozen races in just nine months.

How? The 2.2-liter Nissan NAPS-Z engine. A microprocessor continuously adjusts the fuel/air mixture injected into the engine's crossflow hemihead combustion chambers and selects which of each cylinder's twin sparkplugs to fire, a unique concept in engine design.

Combustion in this "fast-burn" system is completed within 90° of crankshaft rotation. Ordinary engines don't complete combustion until 170°, wasting both power and fuel. The Nissan NAPS-Z saves both by burning gas "faster."

Most important, you can have this scientific advance without compromising looks or luxury features. This unique power plant delivers its high performance wrapped in a sculptured body that turns science into art. See and test drive the 200-SX at your Datsun dealer now.

38 EPA EST HWY

27 EPA EST MPG

Use EPA estimated mpg for comparison, with standard 5-speed. Actual mpg may differ depending on speed, trip length and weather. Highway mpg will probably be less.

DATSUN
WE ARE DRIVEN

BELT YOURSELF

Product of NISSAN

THE NAME OF QUALITY
NISSAN
NISSAN MOTORS

RATION IN U.S.A.

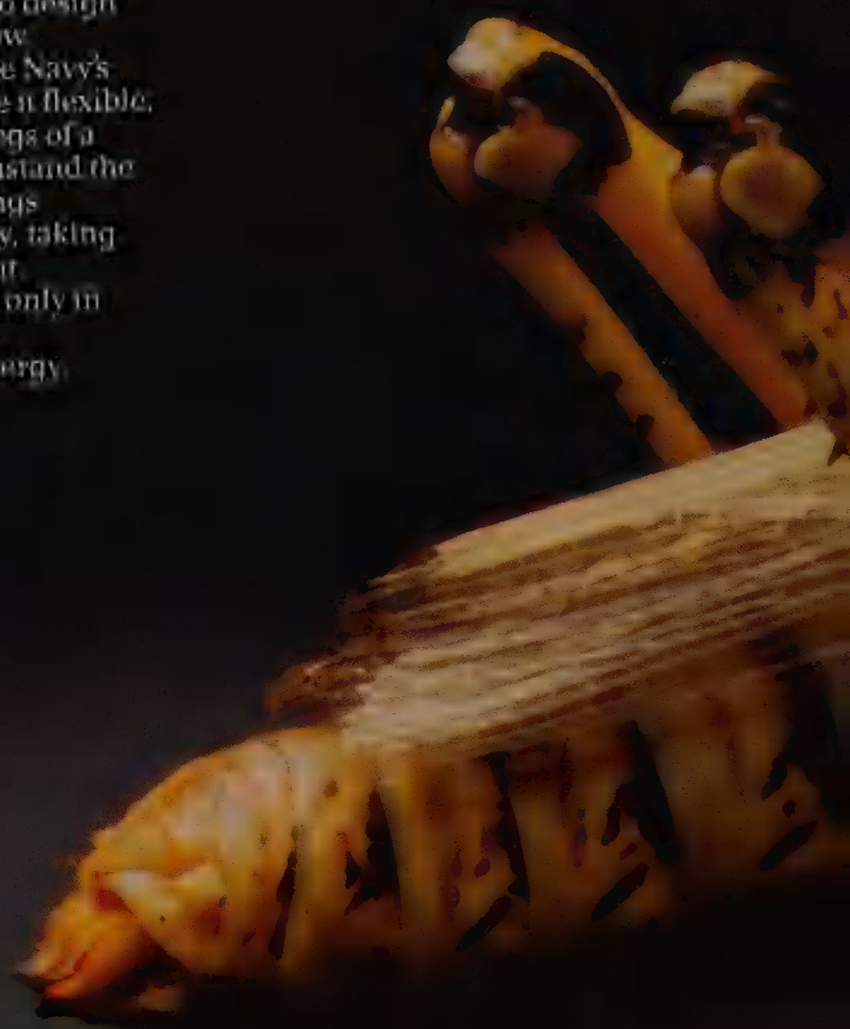
BREAKTHROUGH:
LETS HORNETS LAND LIKE GRASSHOPPERS

When a strike fighter lands on the deck of a carrier, the shock is equal to a fall from a three-story building. The challenge: To design landing gear that will cushion the blow.

When we developed the F/A-18, the Navy's newest fighter-attack aircraft, we gave it flexible, hydraulic "knees" that work like the legs of a grasshopper. They let the Hornet withstand the tough punishment of repeated landings at sea. But they also fold up compactly, taking up little space and adding little weight.

We're creating breakthroughs not only in aerospace but also in health care, information processing, and solar energy.

We're McDonnell Douglas.



MCDONN
DOUGL

ELL
AS

Over 300 manufacturers sell their microcomputers with Microsoft software.

There's a reason.

Powerful simplicity. That's the concept behind every software program we write. Powerful programs that let you spend more time thinking about the problem... and less time thinking about the computer. That concept has earned us the confidence of over 300 microcomputer manufacturers. Starting with the very first microcomputer you could buy. And today, Microsoft® languages, operating systems and applications software are running on well over a million microcomputers. Worldwide.

Made for each other. Manufacturers literally build their computers around Microsoft software. That's no exaggeration. In fact, Microsoft frequently participates in the initial design and development of microcomputers. That's a major reason why Microsoft software runs so well on the majority of the world's 8- and 16-bit systems.

User-oriented. Microsoft is the only software supplier to offer a full range of compatible operating systems, languages, utilities and applications programs. If you're not a computer expert, here's what that means to you: Better programs. Programs that are not only more powerful, but easier to learn and use.

Better tools. We can't buy your trust. We have to earn it. With better products. Tools you can easily use to solve complex problems. We started with the first BASIC for the first microcomputer. Today, we offer a broad range of proven tools for microcomputers. Including SoftCard™ and IBM® RAMCard,™ products that materially increase the capabilities of Apple® and IBM PC computers. Plus, business and management software such as Multiplan,™ the powerful electronic worksheet you can learn to use in just a few hours. Better tools. Because they're designed specifically for your computer.

Ask your Microsoft dealer. Most microcomputer manufacturers offer their systems with some Microsoft software, but you'll undoubtedly want more. Programs and languages that solve your specific problems. Programs for your industry, business or home. Programs that do more, yet ask less. Ask your computer software dealer for a demonstration. You'll see why more than 300 microcomputer manufacturers offer their systems with Microsoft software. The reason is powerful simplicity.

BETTER TOOLS FOR MICROCOMPUTERS

MICROSOFT™

MICROSOFT CORPORATION
10700 NORTHUP WAY
BELLEVUE, WASHINGTON 98004

Microsoft is a registered trademark, and Multiplan, SoftCard, RAMCard and the Microsoft logo are trademarks of Microsoft Corporation. Apple and the Apple logo are registered trademarks of Apple Computer, Inc. IBM and the IBM logo are registered trademarks of International Business Machines Corporation. The Radio Shack logo is a registered trademark of Radio Shack, a Division of Tandy Corporation. The Burroughs logo is a registered trademark of Burroughs Corporation. The Victor logo is a registered trademark of Victor Business Products. The Wang logo is a registered trademark of Wang Laboratories, Inc.

LETTERS

Sirs:

My article "The Galileo Affair" [SCIENTIFIC AMERICAN, August, 1982] has generated a continuing stream of responses, but only recently did Pierre Conway, O.P., remind me that I could have carried the account of the procedures against Galileo one day further. I ended the story on June 21, 1633, the last day on which the panel of cardinals received testimony, and neglected to mention the final sentencing and extended abjuration on June 22, 1633.

In a sense Napoleon is responsible for this lapse. In 1810 the Vatican Archives were hauled to France as part of the aggrandizement of the Bibliothèque Nationale. The records of the Galileo trial were a particularly important desideratum for the Napoleonic officials. When the Archives were returned in 1814, many documents were missing, including this important group. Eventually one of the three record books turned up: the slim volume containing the letters from Cardinal Bellarmine and Galileo's notarization of the testimony, all illustrated in my article. Thus the Vatican Archives today have no 17th-century record of the actual sentencing. A contemporary Italian translation exists in the Modena State Archives (although my request for a copy went unanswered); the Latin version was published in full by the Jesuit Giovanni B. Riccioli in his 1651 *Almagestum novum,* but in writing my article I had temporarily forgotten this source as well as the English translation in Giorgio de Santillana's *The Crime of Galileo.*

In the sentencing the Inquisition accused Galileo of a "vehement suspicion of heresy," and in the loyalty oath that Galileo was forced to read he said in part:

"After it had been notified to me that [this] doctrine was contrary to Holy Scripture, I wrote and printed a book in which I discuss this new doctrine already condemned and adduce arguments of great cogency in its favor without presenting any solution of these. Hence I have been pronounced by the Holy Office to be vehemently suspected of heresy, that is to say, of having held and believed that the Sun is the center of the world and immovable and that the Earth is not the center and moves.

"Therefore, desiring to remove from the minds of your Eminences, and of all Catholic Christians, this vehement suspicion justly conceived against me, with sincere heart and unfeigned faith I abjure, curse and detest the aforesaid errors and heresies and generally every other error and sect whatsoever contrary to the Holy Church, and I swear that in the future I will never again say or assert, verbally or in writing, anything that might furnish occasion for a similar suspicion regarding me; but, should I know any heretic or person suspected of heresy, I will denounce him to this Holy Office...."

From that moment onward the heliocentric doctrine could have been interpreted as heresy de facto by virtue of this legal ruling, in spite of the fact that no official decree had been issued; in this sense Galileo's trial could be considered a trial for heresy even though up until June 22, 1633, no public declaration to that effect had been made. Nevertheless, when Riccioli summed up the arguments 18 years after the event, he wrote that the doctrine that the earth moves "had been condemned as heresy, *or at least as erroneous*" (my italics), thereby expressing the ambiguity; meanwhile, in editions of the Index of prohibited books published in Spain, Copernicus' book was explicitly permitted....

My conclusions therefore stand that at least up until Galileo's sentencing the heliocentric doctrine had not been declared heretical, that he was tried more for disobeying orders than for heresy and that the Vatican today has a curiously restricted set of options in exonerating Galileo.

OWEN GINGERICH

Center for Astrophysics
Harvard College Observatory
Smithsonian Astrophysical
Observatory
Cambridge, Mass.

Sirs:

As an addendum to the fine article by Carl Pomerance, "The Search For Prime Numbers" [SCIENTIFIC AMERICAN, December, 1982], the largest prime number known ($2^{44,497} - 1$) has been superseded by a larger one. The new prime ($2^{86,243} - 1$) was discovered in September, 1982, by David Slowinski, working with a Cray Research Cray-1 computer system. It was confirmed by us, working with a Control Data Corporation Cyber 205 computer. We found this new prime to be a "plus" Mersenne prime.

The confirmation is particularly important because it was made with a new transform multiplication technique that squares numbers in $P \log P$ time. Conventional multiplication approaches require P^2 time (where P is the P in $2^P - 1$). The speed obtained with this technique should enable prime searches in number ranges previously beyond the reach of existing computer technology.

STEPHEN K. MCGROGAN

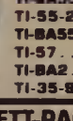
CURT NOLL

Hayward, Calif.

SAVE AT ELEK-TEK ON CALCULATORS

TEXAS INSTRUMENTS

TI-59	170	TI-BA35 (New)	26
TI-58C	80	TI-55-2	36
58/59 Lib.	25	TI-BA55 (New)	43
TI-PC-100C	150	TI-57	29
TI-PROG	50	TI-BA2	36
TI-MBA	30	TI-35-SP	18



NEW! HEWLETT-PACKARD
SLIMLINE Shirtpocket Styled
Power Packed Programmable
LCD PROBLEM SOLVERS

HP-10C	Scientific (New)	$54
HP-11C	Scientific	70
HP-15C	Scientific (New)	92
HP-12C	Financial	92
HP-16C	Programmer (New)	92
HP-97	Desktop Scientific	560

HP-41C	$149
HP-41CV	209
Optical Wand	95
Card Reader	149
Printer	283
Quad R.A.M. (for HP41C)	60

NEW FROM HEWLETT-PACKARD

	Mfr Sugg	Elek-Tek
HP-IL Loop Module	125	95
Digital Cassette Drive	450	343
Printer/Plotter (HP-IL)	450	349
Video Interface	225	175
Ext. Memory Module.	75	60
Ext. Function Module	75	60
Time Module	75	60

Special! 50% off mfr. suggested retail price ($150.00 ea)

HP-34C Scientific	75
HP-38C Financial	75

HUGE SAVINGS on POCKET COMPUTERS from

New! SHARP

PC-1500 POCKET COMPUTER
Extended BASIC system capable of 2 dimensional arrays. Variable string lengths. Program chaining. Full graphic commands, 26 digit LCD
2.6K bytes RAM (Expandable to 10.6K)
16Kbytes of ROM $160
CE-150 Color printer w/cassette interface 165
CE-155 8K Memory 105
CE-158 RS-232 175
CE-159 8-K Memory w/battery backup 125

PC-1211 POCKET COMPUTER
Computing power in mathematics, scientific, engineering & business calculations. Programmable with BASIC language. 24 digit LCD $65
CE-122 16 digit printer w/cassette interface 45

CALL TOLL FREE 800-621-1269
EXCEPT Illinois, Alaska, Hawaii

Access discounted too. Corporate Accts. Inst'l. Mastercard or Visa by mail or phone. Mail Cashier's Check, Money Ord. Pers Check (2wks to clr.) Add $4.00 1st item. (AK, HI, P.R., Canada add $8.00 first item) $1.00 ea. add'l shpg. & handl. Shipments to IL address add 6% tax. Prices subj. to change. WRITE for free catalog. ALL ELEK-TEK MERCHANDISE IS BRAND NEW, FIRST QUALITY AND COMPLETE.

ELEK-TEK, inc.
4557 N. Lincoln Ave., Chicago IL 60625
(800) 621-1269 (312) 477-7840

SCIENCE/SCOPE

The first two satellites to be launched from NASA's space shuttle were placed in orbit for about one-third the cost of a conventional launch, thus saving their owners millions of dollars. This particular model of communications satellite, designed by Hughes Aircraft Company, is relatively inexpensive to launch because it sits upright and snugly folded in special cradles in the cargo bay. This feature saves money because launch costs are based on how much room a satellite takes up and how much it weighs. The cradle contains spring mechanisms that eject the satellite from the bay, after which rocket motors propel it into geostationary orbit. The drum-shaped spacecraft stands 9 feet tall when compacted. But when it reaches orbit, a telescoping solar panel deploys and the antenna unfolds, bringing the satellite's overall height to more than 21 feet.

A compact liquid-crystal light valve is designed to serve as a real-time light modulator for many optical data-processing and projection uses. The Hughes light valve uses liquid-crystal and thin-film technology to combine high input-light sensitivity and high image resolution with low voltage and power requirements. Uses include: graphics projection systems for large-screen displays, high-resolution vision for industrial robots, radar and sonar signal-processing, identification of moving objects, high-resolution spectral analysis of wide-band signals, and hybrid optical-digital processing systems.

Hughes spends about $8 million a year in support of colleges and universities, including help to alleviate the shortage of engineering faculty and modern laboratory equipment. For example, Hughes hires faculty consultants part-time during the school year and full-time in the summer. The program establishes a continuous technical interaction and supplements the pay of faculty members. Also, Hughes donates textbooks, software packages, and a wide variety of used equipment -- including electronic instrumentation and components.

The Smithsonian Institution is installing a new security system to monitor many facilities continuously. The Hughes system includes burglar alarms, fire-sensing devices, voice communications channels, and closed-circuit TV. It will let Smithsonian personnel control entrances and exits, and watch over areas open to visitors. A computer will collect and display information on TV monitors and printers at a central control station. Hughes previously installed a facilities management system at the Smithsonian's National Air and Space Museum. That system provides a wide range of exhibit monitor and control functions.

Hughes is seeking engineers to develop advanced systems and components for many different weather and communications satellites, plus the Galileo Jupiter Probe. Immediate openings exist in applications software development, data processing, digital subsystems test, microwave/RF circuit design, power supply design, digital communications, signal processing, spacecraft antenna design, system integration test and evaluation, and TELCO interconnection. Send your resume to Ray Bevacqua, Hughes Space & Communications Group, Dept. SE, Bldg. S/41, M.S. A300, P.O. Box 92191, Los Angeles, CA 90009. Equal opportunity employer.

Creating a new world with electronics

HUGHES

HUGHES AIRCRAFT COMPANY
EL SEGUNDO, CALIFORNIA 90245

50 AND 100 YEARS AGO

SCIENTIFIC AMERICAN

MARCH, 1933: "There is no question that the great American automobile industry has been hard hit by the general depression. The production of cars declined from 5,600,000 in 1929 to 3,500,000 in 1930, to 2,500,000 in 1931 and finally to fewer than 1,500,000 last year. Consequently, reckoning by the Department of Commerce's estimate of an average length of seven years for all cars, there should be an accumulated replacement demand for more cars in 1933 than were built in 1929. One noted, far-sighted automobile maker says he believes the point of saturation in per capita ownership of cars has not yet been reached. This, together with the present replacement demand and the fact that the foreign market is being more tightly closed against American cars, points to the necessity for concentration on the local market. By concentration, however, is not meant an efflorescence of ballyhoo and high-pressure salesmanship but the production of real and outstanding improvements."

"The next step in aircraft design—the conquest and exploitation of the so-called stratosphere—is close at hand. The two thrilling balloon flights of Professor Auguste Piccard and his aide into the stratosphere to make observations have done much to pave the way for power flights in that region. Once the problems of stratosphere flight are tackled, engineers and scientists will probably find a stratum where air travel will always be a pleasure because of the absence of 'bumps' and other atmospheric disturbances. This is the reason our airline operators will follow closely the exploration of the higher altitudes. The high-altitude ship may solve many present-day difficulties in connection with long distance non-stop flights on a profitable commercial basis. The expected higher speeds in the thin air, with sea-level horsepower and fuel load, will leave room for the pay load necessary to make transport operations worth while at lower rates and will make possible twice the present range of from 500 to 700 miles."

"Henry Norris Russell of Princeton University, whose monthly astronomical articles in this magazine are well known to our readers, has just been elected to the presidency of the Ameri-

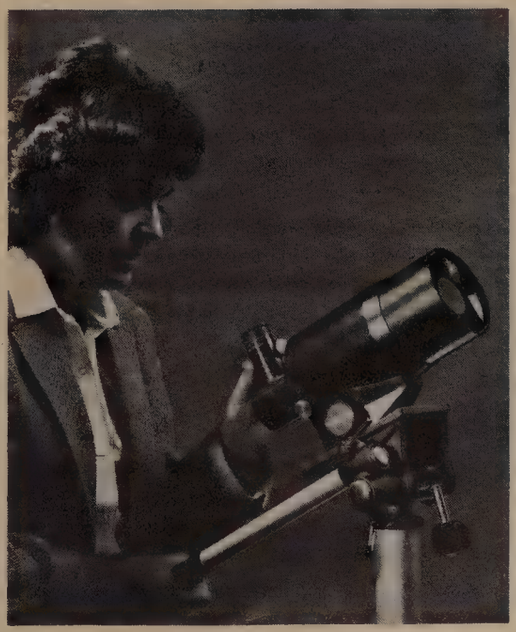
An example of the interesting things that might turn up unexpectedly while you are casually sweeping the deep sky is this supernova in NGC 6946, in a portion of a photograph taken by Hubert Entrop on November 4, 1980, with his Questar. This was 7 days after the official discovery of the nova, as described in Sky & Telescope *in the January, 1981, issue.*

Great news from Questar. . . The Wide-Sky Telescopes

The comet is coming. And Questar is ready for it with two brand-new instruments for all dedicated comet watchers and other deep-sky enthusiasts.

Our emphasis in developing these new designs was on low magnification and optimum field of view—magnification low enough and field of view wide enough to sweep the sky for richfield observing without need of finder or equatorial mount.

We also wanted to dispel the popular misconception that short focal lengths (f/3 to f/5) are essential for effective deep-sky observing and optimum image brightness. Image brightness is a function of aperture and magnification, and has nothing to do with focal length. We continue to use long focal ratios, even in our new wide-sky models. This rules out the comatic aberrations of short-focal-length reflectors and the spurious chromatic effects of short-focal-length refractors, as well as the below-par imagery of short-focal-length eyepieces. Imagine, then, the pleasure of having that familiar sharp Questar® resolution, with no aberrations, available in a telescope with a field of view of 3°.



Specifications of Questar Wide-Sky Telescopes

Questar Wide-Sky 3½ (shown left): aperture 89mm; focal length 700mm; focal ratio f/7.8; magnification 22 × with 32mm wide-angle eyepiece; field of view 3°; faintest visible star 12th magnitude (visual), 12.8 with special coatings.

Questar Wide-Sky 7: aperture 178mm; focal length 2400mm; focal ratio f/13.5; magnification 44 × with 55mm wide-angle eyepiece, 75 × with 32mm wide-angle eyepiece, field of view 1.15° with 55mm eyepiece, 1° with 32mm wide-angle eyepiece; faintest visible star 13th magnitude (visual) with standard coatings, 14th magnitude (visual) with special coatings.

Both instruments can be equipped with various accessories to provide for tracking and photography.

Special mounts for the Wide-Sky Questars are comet seeking and general sky scanning, monitoring telescopic meteors, observing the occultation of stars and planets by the Moon, observing lunar and solar eclipses, observing the deep-sky phenomena — nebulae, star clusters and galaxies — and variable star observing.

Photographic accessories are available for use in these applications, in which the Questar becomes an astrographic camera. All Questars, of course, can be used with a video monitor.

Let us send you our price list.

© 1983 Questar Corporation

QUESTAR

Box C, Dept. 205, New Hope, PA 18938
(215) 862-5277

QUESTAR, THE WORLD'S FINEST, MOST VERSATILE TELESCOPE, IS DESCRIBED IN OUR BOOKLET IN COLOR, WITH PHOTOGRAPHS BY QUESTAR OWNERS, SEND $2 TO COVER MAILING COSTS ON THIS CONTINENT, BY AIR TO SO. AMERICA, $3.50; EUROPE AND NO. AFRICA, $4; ELSEWHERE, $4.50.

Speak French like a diplomat!

What sort of people need to learn a foreign language as quickly and effectively as possible? *Foreign service personnel,* that's who.

Now you can learn to speak French just as these diplomatic personnel do— with the Foreign Service Institute's Basic French Course.

The U.S. Department of State has spent thousands of dollars developing this course. It's by far *the most effective way* to learn French at your own convenience and at your own pace.

The Basic French Course consists of a series of cassettes and an accompanying textbook. Simply follow the spoken and written instructions, listening and repeating. By the end of the course, you'll be learning and speaking entirely in French!

This course turns your cassette player into a "teaching machine." With its unique "pattern drill" learning method, you set your own pace — testing yourself, correcting errors, reinforcing accurate responses.

The FSI's Introductory Basic French Course comes in two parts, each shipped in a handsome library binder. Part A introduces the simpler forms of the language and a basic vocabulary. Part B presents more complex structures and additional vocabulary. You may order one or both:

☐ **Basic French, Part A.** 12 cassettes (17 hr.) and 200-p. text, $125.
☐ **Basic French, Part B.** 18 cassettes (25½ hr.), and 300-p. text, $149.

(Conn. and N.Y. residents add sales tax.)

TO ORDER BY PHONE, PLEASE CALL TOLL-FREE NUMBER: **1-800-243-1234.**

To order by mail, clip this ad and send with your name and address, and a check or money order — or charge to your credit card (AmEx, VISA, MasterCard, Diners) by enclosing card number, expiration date, and your signature.

The Foreign Service Institute's French course is unconditionally guaranteed. Try it for three weeks. If you're not convinced it's the fastest, easiest, most painless way to learn French, return it and we'll refund every penny you paid. Order today!

81 courses in 26 other languages also available. Write us for free catalog. Our 10th year.

Audio-Forum
Suite 33 E
On-the-Green,
Guilford, CT 06437
(203) 453-9794

AUDIO-FORUM®

can Association for the Advancement of Science, the largest of all organized bodies of general scientific men. The A.A.A.S. has almost 20,000 members and comprises practically the entire personnel of the many branches of American science—mathematics, astronomy, physics, chemistry, geology, zoology and so on. It is unofficially the great general 'holding company,' federation or overhead organization of the several sciences, each of which has its own organization, and the award of its presidency is regarded by all men of science as a signal honor."

SCIENTIFIC AMERICAN

MARCH, 1883: "Last year 37,951 persons died in New York City, the ratio being a little over 29 per 1,000 of population. The number of deaths from the principal contagious diseases was as follows: smallpox, 269; measles, 912; scarlet fever, 2,070; diphtheria, 1,521; croup, 730; erysipelas, 151; typhus fever, 66; typhoid fever, 363; malarial fever, 533."

"The relations existing in matter during combustion under atmospheric pressure extend to those creative fires whose ashes form new worlds. On this hypothesis we are led to a chemical condition or period when compounds can begin to form. The combination would necessarily take place in the outer and cooler portions of a star's atmosphere. From this behavior of first principles or elements are derived not only all the substances—solids, liquids and gases—of the earth but also of all other forms of matter throughout the universe. Such are the world fires in the nebular theory of creation. We are, however, led back beyond this point into a gaseous chaos when the whole universe, inconceivable ages ago, was equally filled with a homogeneous mass of tenuous matter at an extremely high degree of temperature. The millions of bodies composing the different solar systems originated only in consequence of rotary movement during which a number of masses acquired greater density than the remaining gaseous mass and then acted upon the latter as central points of attraction."

"Consul E. L. Baker of Buenos Ayres sets forth the success that has attended the introduction of *Eucalyptus globulus* in the Argentine Republic. He has several times in his annual reports referred to the successful introduction of *Eucalyptus* (the blue gum tree of Australia) into the Argentine, and has spoken of the rapidity with which some portions of the pampas heretofore destitute of timber are now being dotted with plantations of these magnificent trees; from the ease with which they can be grown, and the marvelous rapidity of their growth, he has suggested the feasibility of their cultivation in the milder parts of the United States. From what Consul Baker has observed during his stay in the Argentine Republic he is more and more convinced that the eucalyptus is a most desirable tree with which to timber our Southwestern plains and renew our rapidly decreasing forests."

"Professor D. E. Hughes, F.R.S., has formed a theory of magnetism entirely based upon experimental results, and these have led him to the following conclusions: 1. Each molecule of a piece of iron, steel or magnetic metal is a separate and independent magnet, having its two poles and distribution of magnetic polarity exactly the same as its total evident magnetism when noticed upon a steel bar magnet. 2. Each molecule, or its polarity, can be rotated in either direction upon its axis by torsion, stress or physical forces such as magnetism and electricity. 3. The inherent polarity or magnetism of each molecule is a constant quantity like gravity; it can be neither augmented nor destroyed. 4. When we have external neutrality, or no apparent magnetism, the molecules or their polarities arrange themselves so as to satisfy their mutual attraction by the shortest path and thus form a complete closed circuit of attraction. 5. When magnetism becomes evident, the molecules or their polarities have all rotated symmetrically in a given direction, producing a North Pole if rotated in this direction, as regards the piece of steel, or a South Pole if rotated in the opposite direction."

"The *Journal des Usines à Gaz,* on the subject of the use of ammoniacal liquor as a manure, states that it is so highly appreciated by Belgian agriculturists that the entire production of the gas works at Malines was bought up. Upon newly cleared ground the liquor was used just as it left the works, but for irrigation purposes it was diluted with three or four times its bulk of water. The effect produced on the soil by the use of the liquor is stated to be exactly the same as when stable dung is employed."

"The Coast and Geodetic Survey steamer *Blake* returned to New York on February 14 from a winter cruise for deep-sea exploration between the Bermudas and the Bahamas. On January 19, in latitude 19°41′ N, longitude 66°24′ W, about 105 miles northwest of St. Thomas, there was found the greatest depth ever measured in the Atlantic, or 4,561 fathoms. The place was about 80 miles southwest of the place where the *Challenger* made her deepest sounding of 3,862 fathoms."

4

Fourth in a series on how Delco Electronics and Bose technology contribute to your enjoyment of driving.

"...you have to hear it to believe it!"

Popular Mechanics

"A new and revolutionary sound system has been developed by the most unlikely partnership ever to be created in the audio industry...General Motors...has teamed up with Bose Corporation to create a car stereo system that is so far ahead of anything currently available in car audio that audio enthusiasts who can afford it may well be spending more time listening to music in their cars than they do at home."

Len Feldman, Audio Times

"The performance of the Delco-GM/Bose Music System was astounding...I can't imagine anyone...buying (one of these cars) without the Music System."

Gary Stock, High Fidelity

"If your car is this well equipped, you won't want to go home again."

Rich Warren, Chicago Magazine

The Delco-GM/Bose Music System is available as a factory installed option on Cadillac Seville and Eldorado, Buick Riviera, and Oldsmobile Toronado.

Sound so real it will change how you feel about driving.

Delco GM BOSE

709-XNY

The Mercedes-Benz 300 SD Turbodiesel is one of the most expensive, least extravagant corporate automobiles you can buy.

The 300 SD is $38,000* worth of automotive enlightenment–a corporate flagship meant to stress not pomp but efficiency. Meanwhile, its retained value over the past three years has been shown to *average* 90 percent.

Does your company have the technology of the 1990's in its executive offices, and the technology of the 1950's in its executive garage?

The fact is that inefficient operation, excessive size and savage depreciation need *not* define today's corporate flagship. The 300 SD Turbodiesel represents a different and refreshingly more sane equation.

10,000 miles @ $380?

The 300 SD's basic design efficiencies permit the efficiency of a diesel engine only *three liters* in size. It is no ponderous V-8 but an in-line five–and such a light drinker that a fuel cost of about $380 per annum is conceivable, based on 10,000 miles' driving, current average diesel fuel prices of $1.25 per gallon, and a 33 mpg EPA highway mileage figure. (City mpg 27 EPA.)**

The dollars and cents may not be crucial; the concept of relentless efficiency that such figures reflect, however, is as bracing as it is novel in the world of company cars.

So sanguine is Mercedes-Benz about the reliability of the 300 SD and its engine, incidentally, that it comes with a 36-month-or-36,000-mile warranty.†

This frugal diesel is meanwhile the most *powerful* such engine yet placed in a production automobile. Old images of "diesel lag" are obliterated in a turbinelike rush of energy and brisk acceleration, generated by a built-in turbocharger.

Five adults will find themselves extremely well cared for aboard the 300 SD. Its interior is almost 109 cubic feet worth of first-class repose, outfitted with only first-class amenities.

Yet so disciplined is its design that the car is less than 17 feet long and turns within just 39 feet. At 3,780 pounds, it is neither ponderous nor flimsy but athletically trim.

Its all-welded body achieves remarkable solidity and strength in part by using high strength, low alloy steels. Both the trunk lid and hood are fabricated in aluminum–part of the engineers' relentless quest to pare off weight wherever possible.

Not for boulevards only

If the 300 SD doesn't perform like the usual diesel, neither does it handle like the usual limousine in the clutch.

Mercedes-Benz, refreshingly, assumes that even company presidents must sometimes face switchback curves, slippery spots and potholes. The 300 SD is girded for such adversity: its fully independent suspension, zero-offset front suspension geometry and forged light-alloy wheels may be less showy than opera windows or carriage lamps.

There is no real need for a company driver to pilot the 300 SD. Its acute precision of response makes it deeply pleasurable to drive, across town or across country. A "driver's car," defined. For instance, note that its crisp four-speed automatic gearbox can also be shifted by hand.

120 safety features

Passengers are hardly ignored. Twin reading lamps are recessed in the rear of the cabin. There is even a separate ventilation console to serve the rear-seat occupants. The complement of standard features includes electronic cruise control, electric window lifts and front-seat adjustment, AM/FM stereo radio/cassette player with four speakers, and trimming in genuine hand-finished woods. No fewer than 120 safety features are also standard.

The 300 SD's history of value retention outstrips that of any luxury sedan extant. In fact, the *N.A.D.A. Official Used Car Guides* for 1982 calculate an average retained value for the 300 SD, over the past three years, of 90 percent. This helps place its $38,000 price in proper perspective.

A persuasive statement

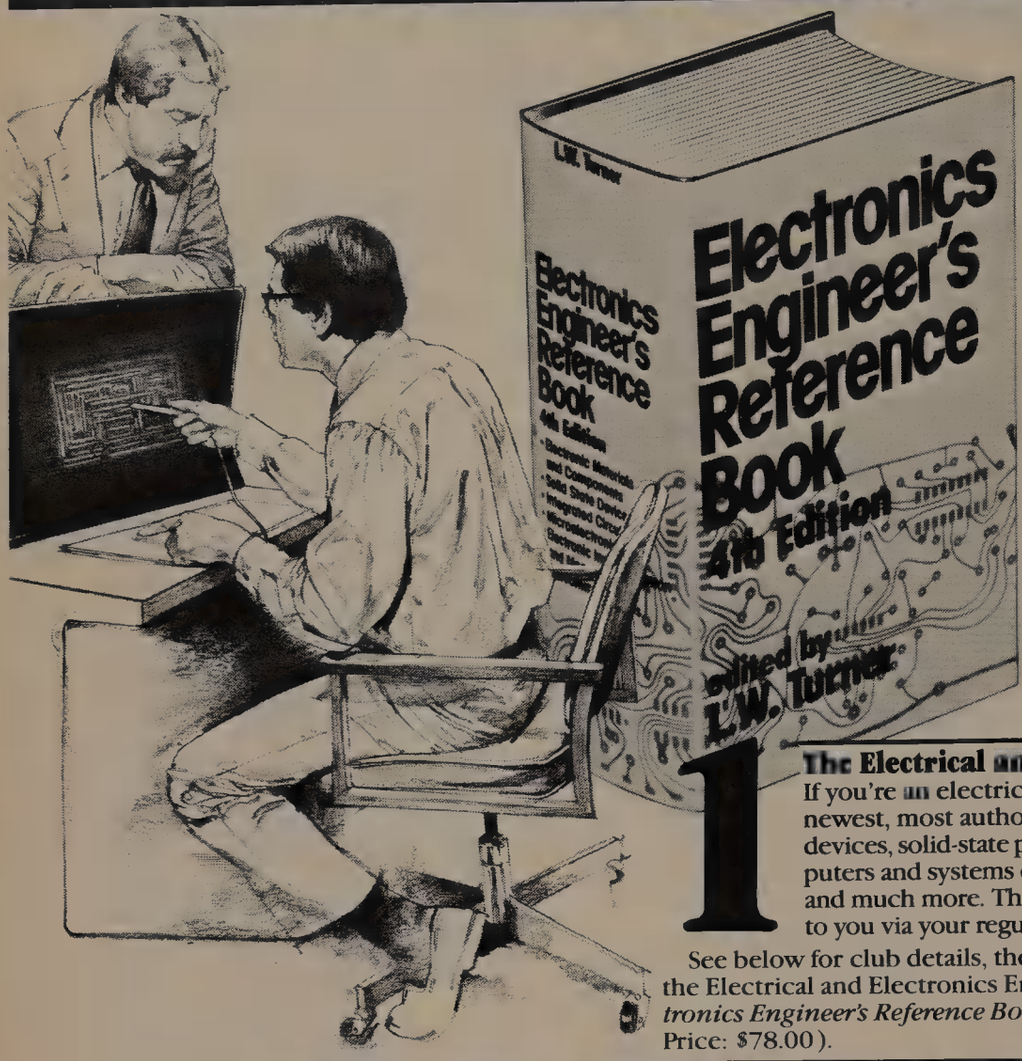
There is one final argument for bringing a Mercedes-Benz 300 SD Turbodiesel Sedan into the firm. While efficiently serving the firm, it serves also as a fine advertisement–identifying your company as progressive-minded, and sensible, and shrewd. What limousine has ever made a statement quite like that?

*Approximate suggested advertised delivered price at port of entry. **EPA estimate for comparison purposes. The mileage you get may vary with trip length, speed and weather. †This is, of course, a limited warranty; you should consult your authorized Mercedes-Benz dealer for full details.

©1983 Mercedes-Benz N.A., Inc., Montvale, N.J.

Engineered like no other car in the world

Your Choice – One of Macmillan's



Pub. Ed. Price $78.00

Only $2.95

when you join the

ELECTRICAL AND ELECTRONICS ENGINEERING BOOK CLUB

- 4th Edition—totally revised and updated
- Over 1,600 pages
- 27 all-inclusive sections
- 61 renowned contributors
- Thousands of illustrations
- Complete index and bibliography
- Full treatment of subjects like Integrated Circuits, Solid-State and Photo-Electronic Devices, Micro-Electronics, Telecommunications, Sound and Video Recording, Electronic Materials and Components, Electronics in Industry, and much more.

1

The Electrical and Electronics Engineering Book Club

If you're an electrical or electronics engineer, you can get the newest, most authoritative books on such topics as microwave devices, solid-state power supplies, integrated electronics, computers and systems design, data communications networks, and much more. These hard to find books will be made available to you via your regular club bulletin at up to 30% savings.

See below for club details, then use the card bound into this ad to join the Electrical and Electronics Engineering Book Club and get your *Electronics Engineer's Reference Book, 4th Edition* for only *$2.95* (Publisher's Price: $78.00).

Pub. Ed. Price $84.00

Only $2.95

when you join the

PROFESSIONAL CIVIL ENGINEERING BOOK CLUB

- 3rd Edition—completely updated, particularly in areas where there is a paucity of published material
- Over 1,800 pages
- Thousands of figures, charts, tables and schematics
- Full treatment of important subject areas such as Materials and Structures, Loading, Hydraulics, Site Investigation, Soil and Rock Mechanics, Concrete and Steel Design, Demolition, Surveying and Photogrammetry, Sewage Disposal, River Maintenance, Heavy Welded Structural Fabrication, and much more.

2

The Professional Civil Engineering Book Club

Make sure that you keep up with latest theories and newest techniques in civil engineering. Our Editors will keep you up to date on areas like construction management and organization, concrete and steel design, hydraulics, all aspects of structural design and much more. The most important books will be readily available to you at low club prices.

See below for club details, then use the card bound into this ad to join the Professional Civil Engineering Book Club and get your *Civil Engineer's Reference Book, 3rd Edition* for only *$2.95* (Publisher's Price: $84.00).

Professional Engineering Book Clubs

SAVE UP TO $81.00
off Publishers' Prices When You Join with this Offer!

Macmillan has three nationally-known Professional Engineering Book Clubs designed for continuing education and career growth: one for Electrical and Electronics Engineers, one for Civil Engineers and one for Mechanical Engineers. If you're a practicing engineer—rely on Macmillan to help you maintain a competitive edge that allows you to move ahead in your job and excell in your chosen field of engineering.

Pub. Ed. Price $79.95

Only $2.95

when you join the

PROFESSIONAL MECHANICAL ENGINEERING BOOK CLUB

- 11th Edition, revised and expanded
- Over 1,700 pages
- 20 all-inclusive sections
- Thousands of charts, tables and schematics
- Covers important areas such as Units, Symbols and Constants, Theory and Design, Metallurgy and Non-Metals Technology, Welding and Surface Finishes, Non-Destructive Testing, Lubricants, Instrumentation, Plastics, Computers in Design, Thermodynamics, Fluid Mechanics, and much more.

3

The Professional Mechanical Engineering Book Club Now get the books you need to get ahead in mechanical engineering—definitive books that deal with subjects like materials and processes in manufacturing; measurements and control applications; reliability; mini and microcomputer controls in the industrial process; automation; pressure vessel design, and much more.

See below for club details, then use the card bound into this ad to join the Professional Mechanical Engineering Book Club and get your *Mechanical Engineer's Reference Book, 11th Edition* for only *$2.95* (Publisher's Price: $79.95).

FOUR GOOD REASONS TO JOIN

1. The Finest Books. Of the hundreds and hundreds of books submitted to us each year, only the very finest are selected and offered. Moreover, our books are always of equal quality to publishers' editions, *never* economy editions.

2. Big Savings. In addition to getting the **Engineer's Reference Book of your choice** for only $2.95 when you join, you keep saving substantially—up to 30% and occasionally even more. (For example, your total savings as a trial member, including this introductory offer, can easily be over 50%. That's like getting every other book free!)

3. Bonus Books. Also, you will immediately become eligible to participate in our Bonus Book Plan, with savings of up to 70% off the publishers' prices.

4. Convenient Service. At 3-4 week intervals (16 times per year) you will receive the Book Club NEWS, describing the Main Selection and Alternate Selections, together with a dated reply card. If you want the Main Selection, do nothing and it will be sent to you automatically. If you prefer another selection or no book at all, simply indicate your choice on the card, and return it by the date specified. You will have at least 10 days to decide. If, because of late mail delivery of the NEWS, you should receive a book you do not want, we guarantee return postage.

If reply card has been removed, please write to Macmillan Professional Engineering Book Club, Dept.—AK6, Riverside, N.J. 08075 to obtain membership information and an application.

THE AUTHORS

ELISABETH MANN BORGESE ("The Law of the Sea") is professor of political science at Dalhousie University in Nova Scotia. The daughter of the novelist Thomas Mann, she was born in Munich and received a diploma from the Conservatory of Music in Zurich in 1937. In 1938 she emigrated to the U.S. with her family, becoming a naturalized citizen in 1941. Before she joined the faculty at Dalhousie in 1978 she was for many years a senior fellow at the Center for the Study of Democratic Institutions in Santa Barbara. The law and politics of the oceans are her major interests. She is the author of many articles and books on the subject, including *The Ocean Regime* (1968) and *Seafarm: The Story of Aquaculture* (1980).

DAVID A. PATTERSON ("Microprogramming") is associate professor of electrical engineering and computer science at the University of California at Berkeley. Before earning his doctorate in 1976, he worked for the Hughes Aircraft Company designing and evaluating computers. He has been at Berkeley since 1977. He spent the fall of 1979 on leave at the Digital Equipment Corporation developing techniques for the design of microprograms. In one of his courses at Berkeley he and his students designed a microprocessor called RISC I (for Reduced-Instruction-Set Computer I), in which all programs in effect become microprograms.

WALTER A. FAIRSERVIS, JR. ("The Script of the Indus Valley Civilization"), is professor of anthropology at Vassar College. He received an associate of arts degree from the University of Chicago in 1953. He went on to obtain his bachelor's degree at Columbia University and his doctorate in anthropology from Harvard University. His professional career has been divided between teaching and leading anthropological expeditions to the Middle East and Far East. He has led expeditions to Afghanistan, Pakistan, Iran and India, some of them under the auspices of the American Museum of Natural History, of which he is a research associate. He held academic appointments at New York University and the University of Washington before moving to Vassar in 1968. Fairservis writes: "I have an implicit belief that in the origins of civilization we are dealing with something we do not completely understand. My whole career has been aimed at increasing our understanding of that."

LESLIE A. GRIVELL ("Mitochondrial DNA") was born and educated in England, receiving his B.S. in 1966 and his Ph.D. in biochemistry in 1969 from University College London. He writes: "My interest in mitochondrial biogenesis was aroused during my undergraduate studies and I was particularly fortunate to enter the field at the time of its upswing in the late 1960's and early 1970's." Since 1969 he has been at the Laboratory of Biochemistry of the University of Amsterdam, where he is currently professor of macromolecular biology. Outside the laboratory he describes himself as an "enthusiastic although somewhat inexpert musician. The instruments I play (recorders, flute and clarinet) reflect phases of interest that have ranged from the Renaissance through baroque and classical periods to the romantics."

DUANE A. DICUS, JOHN R. LETAW, DORIS C. TEPLITZ and VIGDOR L. TEPLITZ ("The Future of the Universe") are physicists who share an interest in the relations between high-energy physics and cosmology. Dicus is associate professor of physics at the University of Texas at Austin. His bachelor's and master's degrees are from the University of Washington and his doctorate in physics is from the University of California at Los Angeles. After serving as research associate at U.C.L.A., the Massachusetts Institute of Technology and the University of Rochester he moved to the University of Texas in 1973. Letaw is technical director of the Severn Communications Corporation in Severna Park, Md. He received his B.A. at Clark University in 1975 and his Ph.D. in physics from the University of Texas at Austin in 1981. Doris Teplitz is currently visitor at the University of Maryland at College Park. Her B.S. is from Wellesley College and her Ph.D. in physics is from Northeastern University. From 1974 to 1978 she was a member of the faculty of the Virginia Polytechnic Institute and State University. Vigdor Teplitz is deputy chief of the Strategic Affairs Division of the U.S. Arms Control and Disarmament Agency and adjunct professor of physics at the University of Maryland. His B.S. is from M.I.T.; his Ph.D. in physics is from the University of Maryland. Before taking up his current jobs he held appointments at M.I.T. and Virginia Polytechnic Institute and State University. In 1979 he served as adviser to the U.S. delegation at the second and third sessions of the negotiations on antisatellite weapons between the U.S. and the U.S.S.R.

SARA J. SHETTLEWORTH ("Memory in Food-hoarding Birds") is professor of psychology at the University of Toronto. She was graduated from Swarthmore College in 1965 with a B.A. and went on to get her M.A. in 1966 from the University of Pennsylvania and her Ph.D. in psychology in 1970 from the University of Toronto. She writes: "My Ph.D. thesis describes how reward and punishment have different effects on different activities such as grooming, digging and rearing in golden hamsters. More recent interests include the role of learning and memory mechanisms in foraging behavior. I have also done work on sea-turtle behavior with my husband Nicholas Mrosovsky."

IRVING R. EPSTEIN, KENNETH KUSTIN, PATRICK DE KEPPER and MIKLÓS ORBÁN ("Oscillating Chemical Reactions") are chemists with a common interest in the subject of their article. Epstein is professor of chemistry at Brandeis University. He holds three degrees from Harvard: an A.B. from Harvard College and an M.A. and a Ph.D. from Harvard University. He went to Brandeis in 1971. In 1977 and 1978 he was a National Science Foundation fellow at the Max Planck Institute for Biophysical Chemistry in Göttingen. Kustin is also professor of chemistry at Brandeis. He received his B.S. from Queens College in 1955. His Ph.D. in inorganic chemistry is from the University of Minnesota. After serving as U.S. Public Health Service fellow at the Max Planck Institute for Physical Chemistry in Berlin he moved to Brandeis in 1961. From 1974 to 1977 he served as chairman of the department of chemistry there. De Kepper is a manager of research in chemical engineering at the French National Center for Scientific Research (CNRS) Paul Pascal Research Center at Domaine University in Talence. A native of France, he was educated at the University of Bordeaux I, earning his Ph.D. in chemical engineering in 1978 and going on to the CNRS laboratory in the same year. Orbán was born and educated in Hungary. His Ph.D. is from Eötvös University in Budapest, where he has been a member of the faculty since 1962; he is currently associate professor of chemistry.

STEPHEN A. GELLER ("Autopsy") is professor of clinical pathology and vice-chairman of the department of pathology at the Mount Sinai School of Medicine. He obtained his undergraduate education at Brooklyn College and went on to earn his M.D. in 1964 from the College of Medicine of Howard University. He was an intern at Lenox Hill Hospital in New York City and resident in pathology at Mount Sinai. He served for two years as pathologist at the Naval Hospital in Beaufort, S.C., before proceeding to Mount Sinai to take up his current work. His specialty is the diseases of the gastrointestinal tract.

Professional achievements are often born of a natural desire to reach out beyond professional traditions. But reaching out, even in a professionally responsible manner, can create new and unknown liability exposures. Fear of these potentially career-damaging exposures too often results in inaction, rather than achievement.

The CNA Insurance Companies recognize the stifling effect these fears can have on professional performance. For more than a quarter-century, we've been providing liability insurance which can encourage professional achievement by easing those fears.

By seeking close association involvement in every aspect of the program, CNA professional liability specialists are better able to tailor coverages to both current and changing needs. Another part of this interchange is CNA's sharing of loss data and loss expenses with associations to give members a clear understanding of their premium base.

Learn how CNA's professional liability insurance can encourage you and the members of your professional group to reach out to new achievements. Ask your independent insurance agent or your association executive to contact Vice President, Professional Liability, CNA, CNA Plaza, Chicago, IL 60685.

"WHAT WOULD LIFE BE IF WE HAD NO COURAGE TO ATTEMPT ANYTHING?"

Vincent VanGogh

CNA Professional Liability Insurance protects professional group members including ▫ Doctors ▫ Attorneys ▫ Architects ▫ Engineers Accountants ▫ Corporate Directors and Officers ▫ School Boards ▫ Hospitals ▫ Medical Clinics ▫ Publishers ▫ Broadcasters

INSURANCE FROM CNA

METAMAGICAL THEMAS

Tripping the light recursive in Lisp, the language of artificial intelligence

by Douglas R. Hofstadter

Since I ended last month's column with a timely newsbreak about the homely Glazunkian porpuquine, I felt it only fitting to start off this month's with more about that little-known but remarkable beast. As you may remember, the quills on any porpuquine (except the tiniest ones) are smaller porpuquines. The tiniest porpuquines have no quills but do have a nose, and a very important nose at that, since the Glazunkians base their entire monetary system on that little object. Consider the value of three-inch porpuquines in Outer Glazunkia. Each porpuquine always has nine quills (contrasting with its cousins in Inner Glazunkia, which always have seven); thus each has nine two-inch porpuquines sticking out of its body. Each of those in turn sports nine one-inch porpuquines, out of each of which sprout nine zero-inch porpuquines, each of which has one nose. All told this comes to $9 \times 9 \times 9 \times 1$ noses, which means that a three-inch porpuquine in Outer Glazunkia has a buying power of 729 noses. If, in contrast, we had been in Inner Glazunkia and had started with a four-incher, that porpuquine would have a buying power of $7 \times 7 \times 7 \times 7 \times 1$, or 2,401, noses.

Let us see if we can come up with a general recipe for calculating the buying power (measured in noses) of any porpuquine. It seems to me that it would go something like this:

The buying power of a porpuquine with a given quill count and size is:
if its size equals 0, then 1;
otherwise figure out the buying power of a porpuquine with the same quill count but of the next-smaller size and multiply that by the quill count.

We can shorten this recipe by adopting some symbolic notation. First, let "q" stand for the quill count and "s" for the size. Then let "cond" stand for "if" and "t" for "otherwise." Finally, use a kind of condensed algebraic notation in which the English names of operations are placed to the left of their operands, inside parentheses. We get something like this:

```
(buying-power q s) is:
   cond (eq s 0) 1;
        t (times
            q
            (buying-power
                q
                (next-smaller s)))
```

This is an exact translation of the earlier English recipe into a slightly more symbolic form. We can make it a little more compact and symbolic by adopting a couple of new conventions. Let each of the two cases (the case where s equals 0 and the "otherwise" case) be enclosed in parentheses; in general, use parentheses to enclose each logical unit completely. Finally, indicate by the words "def" and "lambda" that this is a definition of a general notion called "buying power" with two variables (quill count q and size s). Now we get:

```
(def buying-power
 (lambda (q s)
  (cond ((eq s 0) 1)
        (t
         (times q
           (buying-power q
             (next-smaller s)))))))
```

We mentioned above that the buying power of a nine-quill, three-inch porpuquine is 729 noses. This could be expressed by saying that (buying-power 9 3) equals 729. Similarly, (buying-power 7 4) equals 2,401.

Now let us get back to Lisp. I had posed a puzzle toward the end of last month's column in which the object was to write a Lisp function that subsumed a family of functions called "square", "cube", "4th-power", "5th-power" and so on. I asked you to come up with one *general* function called "power", having two variables, such that "(power 9 3)" gives 729, "(power 7 4)" gives 2,401 and so on. I had presented a "tower of power," that is, an infinitely tall tower of separate Lisp definitions, one for each power, connecting it to the preceding power. Thus a typical floor in this tower would be:

```
(def 102nd-power
 (lambda (q)
  (times q (101st-power q))))
```

Of course, 101st-power would refer to 100th-power in its definition, and so on, thereby creating a rather long regress back to the "embryonic," or simplest, case. Incidentally, that very simplest case, rather than "square" or even "1st-power", is this:

```
(def 0th-power (lambda (q) 1))
```

I told you that you had all the information necessary to assemble the proper definition. All you needed to observe is, of course, that each floor of the tower rests on the "next-smaller" floor (except for the bottom floor, which is a "stand-alone" floor). By "next-smaller" I mean the following:

```
(def next-smaller
 (lambda (s) (difference s 1)))
```

Thus "(next-smaller 102)" yields 101. Actually Lisp has a standard name for this, namely "sub1", and a name for its inverse as well, namely "add1". If we put all our observations together, we come up with the following universal definition:

```
(def power
 (lambda (q s)
  (cond ((eq s 0) 1)
        (t
         (times q
           (power q
             (next-smaller s)))))))
```

This is the answer to the puzzle I posed. H'm, that's funny. I have the strangest sense of déjà vu. I wonder why?

The definition presented here is called a *recursive* definition, for the reason that inside the definiens the definiendum is used. This is a fancy way of saying that I appear to be defining something in terms of itself, which ought to be considered gauche if not downright circular in anyone's book. To see whether the Lisp genie looks askance on such trickery let us ask it to figure out (power 9 3):

```
—> (power 9 3)
729
—>
```

Well, fancy that! No complaints? No choking? How can the Lisp genie swallow such nonsense?

The best explanation I can give is to

CIMARRON RESPONDS

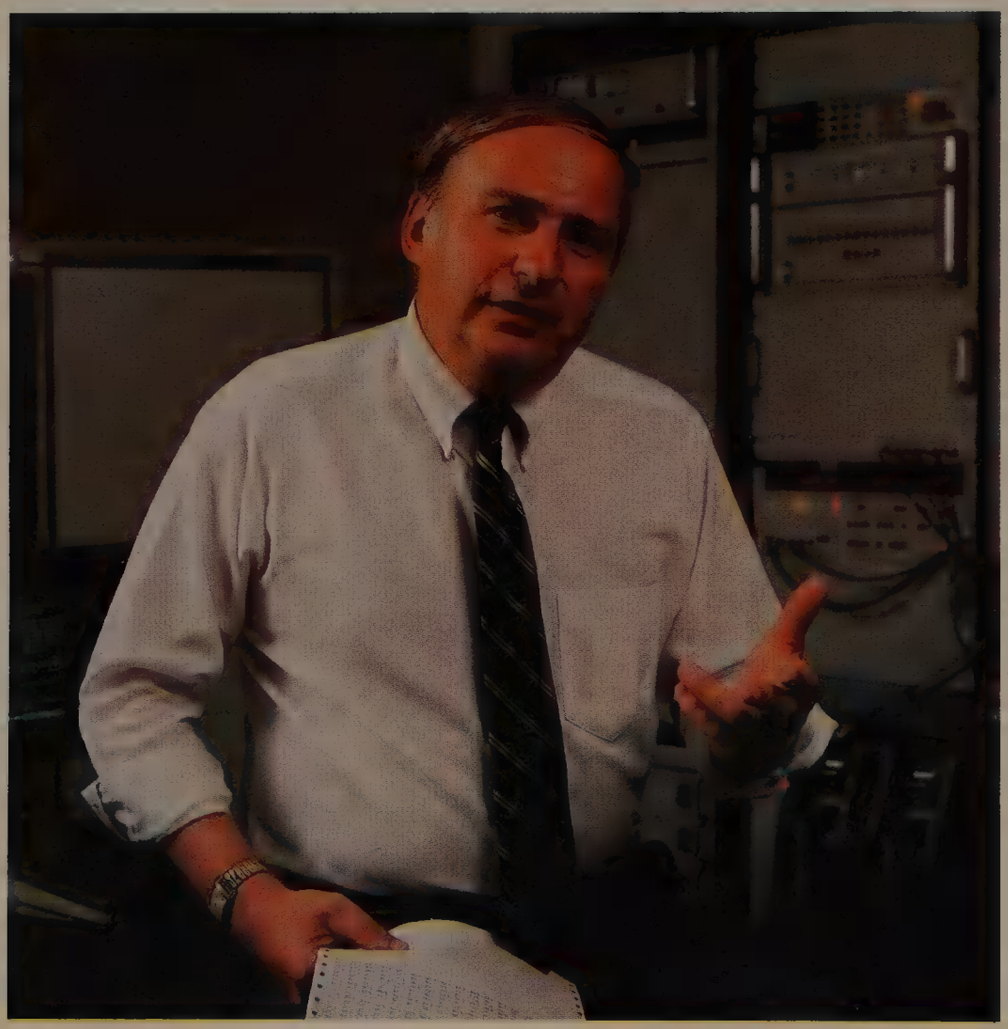
You can sense it the moment you first see Cimarron's bold, contemporary road stance. You can feel it the second you slip behind the wheel and into Cimarron's leather-faced lumbar-supported front bucket seats. And you can experience it every time you take to the open road.

Cimarron responds. See for yourself at your Cadillac dealer's soon. And be sure to ask about the special edition Cimarron D'ORO, shown above.

BEST OF ALL...IT'S A CADILLAC.

GM

What if you chose as a technica



"At M/A-COM, HP computers increase test speed 12-fold."

M/A-COM, Inc., is a major manufacturer of electronic products for commercial telecommunications and defense markets. Dr. Alan Carlson, Manager of the Automati Test Equipment Group in Burlington, Mass., says: "We build automatic test systems and conduct final test on hundreds of M/A-COM products. To automate, we needed precision instruments plus computer power.

"HP was the only vendor to offer us a complete solution. With 20 HP 1000 computers driving more than 200 HP instruments, we've completely automated several product tests. And we can depend on HP's support teams to keep our entire testing network up and running.

"Now we can test sophisticated products 12 times faster than befor With these kinds of savings, each HP 1000 has paid for itself in less than one year."

What should you expect from a computer partner?

Quality—from researching customer needs to product development, manufacturing, marketing, after-sales service, and support!

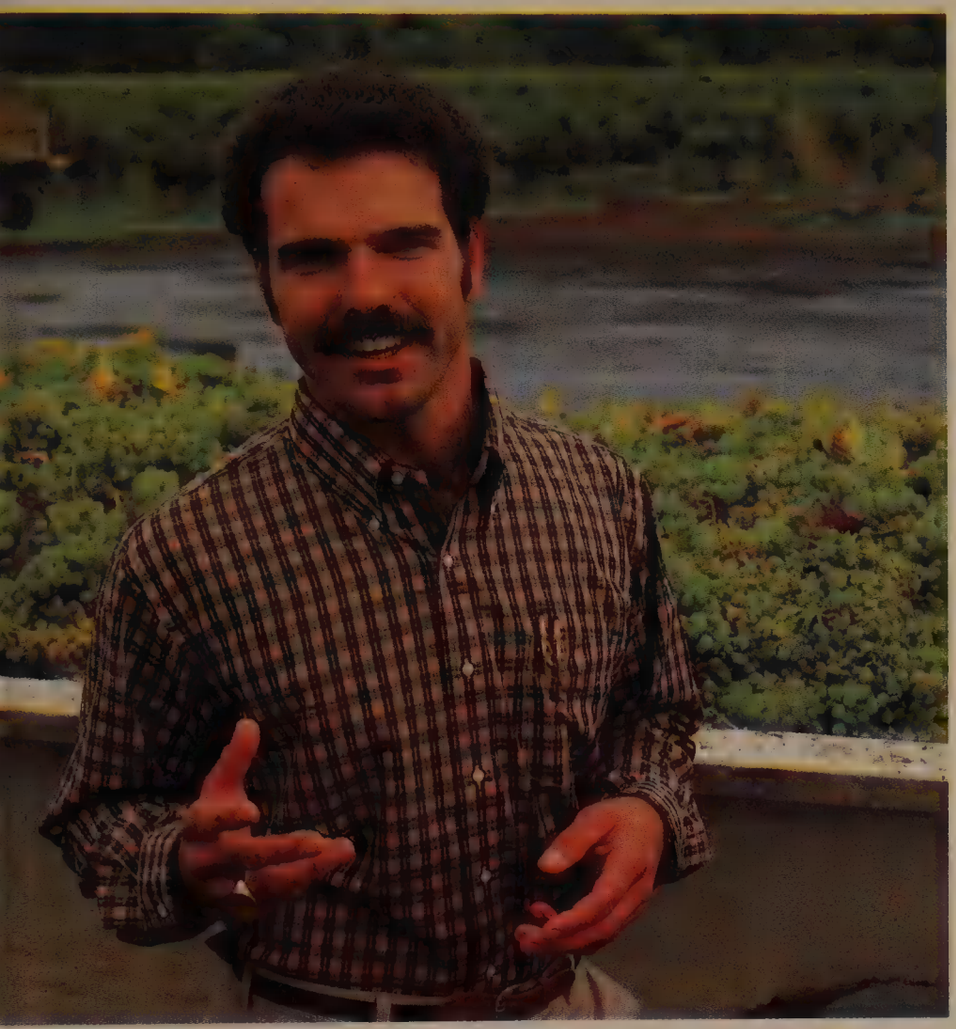
Quality, by HP's definition, encompasses more than the product.

This high quality recently was recognized by the Union of Japanese Scientists & Engineers, who awarded a coveted Deming Prize to Yokogawa-Hewlett-Packard, HP's joint venture company in Japan.

YHP offers the entire spectrum of HP products and in 1981 accounted for Japanese sales of more than $200 million. Nearly 60 percent of the products YHP sells are imported from HP manufacturing operations

The Deming Prize: *for outstanding merit in the control of quality.*

Hewlett-Packard computer partner?



"At Charles Krug Winery, an HP computer tests temperature 150 times faster than manual methods, improves quality control, and doubles production."

Charles Krug and CK Mondavi premium wines are produced at the Charles Krug Winery in St. Helena, California. The winery uses an HP 1000 computer system to monitor and control the temperature of its fermentation tanks.

Treasurer Peter Mondavi, Jr. explains: "The HP 1000 tests the temperature of each tank 150 times faster than manual methods. And because it automatically adjusts fluctuating temperatures we know each bottle of wine will meet our quality standards.

"Improved temperature control has enabled us to double our production while maximizing quality control. As a result, we've saved $50,000 in production costs this year alone."

the US and abroad. Conversely, HP manufactures and exports rious products to HP customers roughout the world.

e total quality control approach which YHP was awarded the ming Prize is a point of pride roughout HP. Customers in every rt of the world should expect *lity* as a fundamental element in omputer partnership with HP.

You should consider a working partnership with HP. Now.

For a free copy of our brochure explaining HP's practical, proven approach to meeting your long-term computer and information needs, write to A.P. Oliverio, Senior Vice President, Marketing, Dept. 207, Hewlett-Packard Company, P.O. Box 10301, Palo Alto, CA 94303.

When performance must be measured by results

HEWLETT PACKARD

0002340

point out that no circularity is actually involved. Although it is true that the definition of "power" uses the word "power" inside itself, the two occurrences are referring to different circumstances. In a nutshell, (power q s) is being defined in terms of a simpler case, namely (power q (next-smaller s)). Thus I am defining the 44th power in terms of the 43rd power, and that in terms of the next-smaller power, and so on down the line until we come to the "bottom line," as I call it: the 0th power, which needs no recursion at all. It suffices to tell the genie that its value is 1, which we did. So when you look carefully, you see that this recursive definition is no more circular than the "tower of power" was—and you cannot get any straighter than an infinite straight line! In fact, this one compact definition really is just a way of getting the entire tower of power into one finite expression. Far from being circular, it is just a handy summary of infinitely many different definitions, all belonging to one family.

In case you still have a trace of skepticism about this sleight of hand, perhaps I should let you watch what the Lisp genie will do if you ask for a "trace" of the function and then ask it once again to evaluate (power 9 3). For this you will need to refer to the illustration below.

On the lines marked "ENTERING" the Lisp genie prints the values of the two arguments, and on the lines marked "EXITING" it prints the value it has computed and is returning. For each ENTERING line there is of course an EXITING line, and the two are aligned vertically, that is, they are indented by the same amount.

You can see that in order to figure out what (power 9 3) is, the genie must first calculate (power 9 2). This, however, is not a given; instead it requires knowing the value of (power 9 1), and this in turn requires (power 9 0). Ah! We *were* given this one: it is just 1. And now we can bounce back "up," remembering that in order to get one answer from the "deeper" answer we must multiply by 9. Hence we get 9, then 81, then 729, and we are done.

I say "we," but of course it is not we but the Lisp genie who must keep track of these things. The Lisp genie has to be able to suspend one computation to work on another one whose answer was asked for by the first one. And the second computation too may ask for the answer to a third one, thus putting itself on hold, as may the third, and so on recursively. Eventually, however, there will come a case where the buck stops—that is, where a process runs to completion and returns a value—and this will enable other stacked-up processes to finally return values, like stacked-up airplanes that have circled for hours finally getting to land, with each landing opening the way for another landing.

Ordinarily the Lisp genie will not print out a trace of what it is thinking unless you ask for it. Whether you ask to see it or not, however, this kind of thing is going on behind the scenes whenever a function call is evaluated. One of the enjoyable things about Lisp is that it can deal with such recursive definitions without getting flustered.

I am not so naive as to expect that you have now totally got the hang of recursion and could go out and write huge recursive programs with ease. Indeed, recursion can be a remarkably subtle means of defining functions, and sometimes even an expert will have trouble figuring out the meaning of a complicated recursive definition. I therefore thought I would give you some practice in working with recursion.

Let me give a simple example based on this riddle: How do you make a pile of 13 stones? A silly answer would be: Put one stone on top of a pile of 12 stones. Suppose we want to make a Lisp function that will give us not a pile of 13 stones but a list consisting of 13 copies of the atom "stone" or, in general, n copies of that atom. We can base our answer on the riddle's silly-seeming yet correct recursive answer. The general notion is to build the answer for n out of the answer for n's predecessor. Build how? Using the list-building function cons, that's how. What is the embryonic case? That is, for which value of n does this riddle present absolutely no problem at all? That's easy: When n equals 0, our list should be empty, which means the answer is nil. We can now put our observations together as follows:

```
(def bunch-of-stones
 (lambda (n)
  (cond ((eq n 0) nil)
        (t
         (cons 'stone
           (bunch-of-stones
             (next-smaller n)))))))
```

Now let us watch the genie put together a very small bunch of stones (with "trace" on, just for fun). For this you will need to refer to the illustration on the opposite page.

This illustrates what is called "consing up a list." Now let us try another one. This one is an old chestnut of Lisp and indeed of recursion in general. Look at the definition and see whether you can figure out what it is supposed to do; then read on to see if you were right.

```
(def wow
 (lambda (n)
  (cond ((eq n 0) 1)
        (t
         (times n (wow (sub1 n)))))))
```

Remember, "sub1" means the same as "next-smaller". For a lark, why don't you calculate the value of (wow 100)? (If you ate your mental Wheaties this morning, try it in your head.)

It happens that Lisp genies often mumble out loud while they are executing wishes, and I just happen to have overheard this one as it was executing the wish "(wow 100)". Its soliloquy ran something like this:

"H'm... (wow 100), eh? Well, 100 surely isn't equal to 0, so I guess the answer has to be 100 times what it *would* have been had the problem been (wow 99). All right—now all I need to do is figure out what (wow 99) is. Oh, this is going to be a piece of cake! Let's see, is 99 equal to 0? No, seems not to be, so I guess the answer to *this* problem must be 99 times what the answer *would* have been had the problem been (wow 98). Let's see..."

At this point the author, having some pressing business at the bank, had to leave the happy genie and did not return until some milliseconds afterward. When he did so, the genie was just finishing up, saying:

"...and now I just need to multiply *that* by 100, and I've got my final answer. Easy as pie! I believe it comes out to be 933262154439441526816992388562667004907159682643816214685929638952175999932299156089414639761565182862536979208272237582511852109168640000000000000000000000000—if I'm not mistaken."

Is that the answer you got, dear reader? No? Oh, I see where you went wrong.

```
—> (power 9 3)
    ENTERING power (q = 9, s = 3)
      ENTERING power (q = 9, s = 2)
        ENTERING power (q = 9, s = 1)
          ENTERING power (q = 9, s = 0)
          EXITING  power (value: 1)
        EXITING  power (value: 9)
      EXITING  power (value: 81)
    EXITING  power (value: 729)
729
—>
```

The Lisp genie evaluates (power 9 3)

It was in your multiplication by 52. Go back and try it again from that point on, and be a little more careful in adding up those long columns. I'm quite sure you'll get it right this time.

This "wow" function is ordinarily called "factorial"; n factorial is usually defined as the product of all the numbers from 1 through n. A recursive definition, however, looks at things a bit differently: speaking recursively, n factorial is simply the product of n and the preceding factorial. It reduces the given problem to a simpler one of the same type. That simpler one will in turn be reduced, and so on down the line, until you come to the simplest problem of that type, which I call the embryonic case or the bottom line. People speak, in fact, of a recursion "bottoming out."

A *New Yorker* cartoon from a few years back illustrates the concept perfectly. It shows a man of 50 or so holding a photograph of himself taken roughly 10 years earlier. In that photograph he is holding a photograph of himself 10 years earlier than *that*. And on it goes, until eventually it bottoms out—quite literally—in a photograph of a bouncy baby boy in his birthday suit (bottom in the air). This idea of recursive photographs catching you as you grow up is quite appealing—I wish my parents had thought of it. Compare it with the more famous infinite regress on the Morton Salt package, where the Morton Salt girl holds a box of Morton Salt with her picture on it. Since the girl in the picture is no younger, however, there is no bottom line and the regress is—at least theoretically—endless.

The recursive approach works when you have a family of related problems at least one of which is so simple that it can be answered immediately. This I call the embryonic case. (In the factorial example that is the "(eq n 0)" case, whose answer is 1.) Each problem (for instance, "What is 100 factorial?") can be viewed as a particular case of one general problem ("How do you calculate factorials?"). Recursion takes advantage of the fact that the answers to various cases are related in some logical way to one another. (For example, I could very easily tell you the value of 100 factorial if only someone would hand me the value of 99 factorial; all I need to do is multiply by 100.) You could say that the recursioneer's motto is "Gee, I could solve *this* case if only someone would magically hand me the answer to the case that is one step closer to the embryonic case." Of course, this motto presumes that certain cases are in some sense "nearer" to the embryonic case than others. In fact, it presumes that there is a natural pathway leading from any case through simpler cases all the way down to the embryonic case, a pathway whose steps are clearly marked.

```
—> (bunch-of-stones 2)
    ENTERING bunch-of-stones (n = 2)
        ENTERING bunch-of-stones (n = 1)
            ENTERING bunch-of-stones (n = 0)
            EXITING    bunch-of-stones (value: nil)
        EXITING    bunch-of-stones (value: (stone))
    EXITING    bunch-of-stones (value: (stone stone))
(stone stone)
—>
```

The genie puts together a very small bunch of stones

As it turns out, this is a most reasonable assumption to make in all kinds of circumstances. To spell out the exact nature of this recursion-guiding pathway you have to answer two Big Questions:

1. What is the embryonic case?
2. What is the relation between a typical case and the next-simpler case?

Now, actually both of these Big Questions break up into two subquestions (as befits any self-respecting recursive question), one concerning how you recognize where you are or how to move, the other concerning what the answer is at any given stage. Thus, spelled out more explicitly, our Big Questions are:

1*a*. How can you know when you have reached the embryonic case?
1*b*. What is the embryonic answer?

2*a*. From a typical case, how do you take exactly one step toward the embryonic case?
2*b*. How do you build this case's answer out of the "magically given" answer to the simpler case?

Question 2*a* concerns the nature of the *descent* toward the embryonic case, or bottom line. Question 2*b* concerns the inverse aspect, namely the *ascent* that carries you back up from the bottom to the top level.

In the case of the factorial the answers to the Big Questions are:

1*a*. The embryonic case occurs when the argument is 0.
1*b*. The embryonic answer is 1.

2*a*. Subtract 1 from the present argument.
2*b*. Multiply the "magic" answer by the present argument.

Notice how the answers to these four questions are all incorporated in the recursive definition of "wow".

Recursion relies on the assumption that sooner or later you will bottom out. One way to be *sure* you will bottom out is to have all the "descending," or simplifying, steps move in the same direction at the same rate, so that your pathway is quite obviously linear. For instance, it is obvious that by subtracting 1 repeatedly, you will eventually reach 0, provided you started with a positive integer. It is also obvious that by performing the list-shortening operation of "cdr" you will eventually reach nil, provided you started with a finite list. For this reason recursions using "sub1" or "cdr" to define their pathway of descent toward the bottom are commonplace. I shall present a cdr-based recursion below, but first I want to show a funny numerical recursion in which the pathway toward the embryonic case is anything but linear and smooth.

Consider the famous "3n + 1" problem, in which you start with any positive integer, and if it is even, you halve it; otherwise you multiply it by 3 and add 1. Let us call the result of this operation on n "(hotpo n)" (standing for "half or triple plus one"). Here is a Lisp definition of hotpo:

```
(def hotpo
 (lambda (n)
  (cond ((even n) (half n))
        (t (add1 (times 3 n))))))
```

This definition presumes that two other functions either have been or will be defined elsewhere for the Lisp genie, namely "even" and "half" ("add1" and "times" being, as I mentioned earlier, intrinsic parts of Lisp). Here are the missing definitions:

```
(def even
 (lambda (n) (eq (remainder n 2) 0)))

(def half (lambda (n) (quotient n 2)))
```

What do you think happens if you begin with some integer and perform hotpo over and over again? Take 7, for instance, as your starting point. Before you do the arithmetic guess at what kind of behavior might result.

As it turns out, the pathway is often surprisingly chaotic and bumpy. For instance, if we begin with 7, the process leads us to 22, then 11, then 34, 17, 52, 26, 13, 40, 20, 10, 5, 16, 8, 4, 2, 1, 4, 2, 1, 4, 2, 1,... Note that we wind up in a short loop. Suppose, therefore, we agree

```
—> (pathway-to-1 3)
ENTERING  pathway-to-1  (tato = 3)
   ENTERING  pathway-to-1  (tato = 10)
      ENTERING  pathway-to-1  (tato = 5)
         ENTERING  pathway-to-1  (tato = 16)
            ENTERING  pathway-to-1  (tato = 8)
               ENTERING  pathway-to-1  (tato = 4)
                  ENTERING  pathway-to-1  (tato = 2)
                     ENTERING  pathway-to-1  (tato = 1)
                     EXITING   pathway-to-1  (value:  (1))
                  EXITING   pathway-to-1  (value:  (2  1))
               EXITING   pathway-to-1  (value:  (4  2  1))
            EXITING   pathway-to-1  (value:  (8  4  2  1))
         EXITING   pathway-to-1  (value:  (16  8  4  2  1))
      EXITING   pathway-to-1  (value:  (5  16  8  4  2  1))
   EXITING   pathway-to-1  (value:  (10  5  16  8  4  2  1))
EXITING   pathway-to-1  (value:  (3  10  5  16  8  4  2  1))
(3 10 5 16 8 4 2 1)
—>
```

How the genie "thinks" when the trace feature is on

that if we ever reach 1, we have hit bottom and can stop. You might well ask, "Who says we *shall* hit 1? Is there a guarantee?" Indeed, if you haven't tried it out a bit, there is no obvious reason to suspect that you will hit 1. (In the case of 7 did you suspect it would happen?) Numerical experimentation, however, reveals a remarkable reliability to the process; it seems that no matter where you start, you always do hit 1 after a while. (Try starting with 27 if you want a real roller-coaster ride!)

Can you write a recursive function to reveal the pathway followed from an arbitrary starting point "down" to 1? Note that I say "down" advisedly, since many of the steps are in fact *up*. Thus the pathway starting at 3 would be the list "(3 10 5 16 8 4 2 1)". In order to solve this puzzle you need to go back and answer for yourself the two Big Questions of Recursion, as they apply here. Note:

```
(cond ((not (want help))
       (not (read further)))
      (t (read further)))
```

First, about the embryonic case. This is easy. It has already been defined as the arrival at 1, and the embryonic, or simplest possible, answer is the list "(1)", a tiny but valid pathway beginning and ending at 1.

Second, about more typical cases. What operation will carry us from typical 7 one step closer to embryonic 1? Certainly not the "sub1" operation. No, by definition it is the function hotpo itself that brings you ever "nearer" to 1—even when it carries you *up*. This teasing quality is of course the entire point of the example. What about 2*b*, how to recursively build a list documenting our wildly oscillating pathway? Well, the pathway belonging to 7 is got by tacking (that is, consing) 7 onto the shorter pathway belonging to (hotpo 7), or 22. After all, 22 is one step closer to being embryonic than 7 is!

These answers enable us to write down the desired function definition, using "tato" as our dummy variable—"tato" being a well-known acronym for "tato (and tato only)", which recursively expands to "tato (and tato only) (and tato (and tato only) only)", and so forth.

```
(def pathway-to-1
 (lambda (tato)
  (cond ((eq tato 1) '(1))
        (t
         (cons tato
               (pathway-to-1
                  (hotpo tato)))))))
```

Look at the way the Lisp genie "thinks" (as revealed when the trace feature is on). Here you should refer to the illustration above.

Notice the total regularity (the *V* shape) of the left margin of the trace diagram, in spite of the chaos of the numbers involved. Not all recursions are so geometrically pretty when they are traced. The reason is that some problems require that *more than one subproblem* be solved. As a practical, real-life example of such a problem, consider how you might go about counting up all the unicorns in Europe. This is certainly a nontrivial undertaking, yet there is an elegant recursive answer: Count up all the unicorns in Portugal (the "car" of Europe, metaphorically speaking), then count up all the unicorns in the other 30-odd countries of Europe (the "cdr" of Europe, to continue the metaphor) and finally add the two results together.

Notice how this spawns two smaller unicorn-counting subproblems, which in turn will spawn two subproblems each, and so on. For instance, how can one count all the unicorns in Portugal? It is easy: add the number of unicorns in the Estremadura region (Portugal's "car") to the number of unicorns in the rest of Portugal (Portugal's "cdr"). And how do you count up the unicorns in Estremadura (not to mention those in the remaining regions of Portugal)? By further breakup, of course. What, then, is the bottom line? Well, regions can be broken up into districts, districts into square kilometers, square kilometers into hectares, hectares into square meters—and we can handle each square meter without further breakup.

Although this may sound rather arduous, there really is no way to conduct a thorough census other than to traverse every single part on every level of the full structure you have, no matter how giant it may be. There is a perfect Lisp counterpart to this unicorn census: it is the problem of determining how many atoms there are inside an arbitrary list. How can we write a Lisp function called "atomcount" that will give us the answer 15 when it is shown the following strange-looking list (which we shall call "brahma")?

```
(((ac ab cb) ac (ba bc ac))
             ab
 ((cb ca ba) cb (ac ab cb)))
```

One method, expressed recursively, is exactly parallel to that for ascertaining the unicorn population of Europe. See if you can write it down.

The idea is this. We want to construct the answer—namely 15—out of the answers to simpler atom-counting problems. Well, it is obvious that one atom-counting problem simpler than (atomcount brahma) is (atomcount (car brahma)). Another is (atomcount (cdr brahma)). The answers to these two problems are respectively 7 and 8. Now, clearly 15 is made out of 7 and 8 by addition—which makes sense, after all, since the total number of atoms must be the number in the car plus the number in the cdr. There is nowhere else for any atoms to hide. This analysis gives us the following recursive definition, with "s" as the dummy variable:

```
(def atomcount
 (lambda (s)
  (plus (atomcount (car s))
        (atomcount (cdr s)))))
```

It looks simple, but it has a couple of flaws. First, we have written the *recursive* part of the definition, but we have utterly forgotten the other equally vital half—the bottom line. It reminds me of a Maryland judge I once read about in the paper, who ruled that "a horse is a four-legged animal that is produced by two other horses." This is a lovely definition, but where does it bottom out? It is the same for atomcount. What is the sim-

plest case, the embryonic case, of atomcount? It is when we are asked to count the atoms in a single atom. The answer in such a case is of course 1. But how can we know when we are looking at an atom? Fortunately Lisp has a built-in function called "atom" that returns t (meaning "true") whenever we are looking at an atom, and returns nil otherwise. Thus "(atom 'plop)" returns t and "(atom '(a b c))" returns nil. Using that, we can patch up our definition:

```
(def atomcount
 (lambda (s)
  (cond ((atom s) 1)
        (t
         (plus (atomcount (car s))
               (atomcount (cdr s)))))))
```

It is still, however, not quite right. If we ask the genie for the atomcount of "(a b c)", instead of getting 3 for an answer we shall get 4. Shocking! How does this happen? Well, we can pin the problem down by trying an even simpler example: if we ask for (atomcount '(a)), we find we get 2 instead of 1. Now the error should be clearer: 2 = 1 + 1, with 1 each coming from the car and the cdr of "(a)". The car is the atom "a", which indeed should be counted as 1, but the cdr is nil, which should not. Then why does nil give an atomcount of 1? Because nil is not only an empty list but also an atom! To suppress this bad effect we simply insert another cond clause at the very top:

```
(def atomcount
 (lambda (s)
  (cond ((null s) 0)
        ((atom s) 1)
        (t
         (plus (atomcount (car s))
               (atomcount (cdr s)))))))
```

I wrote "(null s)", which is just another way of saying "(eq s nil)". In general if you want to determine whether the value of some expression is nil or not, you can use the inbuilt function "null", which returns t if yes and nil if no. Hence, for example, "(null (null nil))" evaluates to nil, since the inner function call evaluates to t, and t is not nil! What about "(null '(null nil))"?

Now, what happens when we run our function on brahma, its original target? The result, with trace on, is shown below.

Notice the more complicated topography of this recursion, with its ins and outs along the left margin. The preceding *V*-shaped recursion looked like a

```
—>(atomcount brahma)
  ENTERING atomcount
    (s = (((ac ab cb) ac (ba bc ac)) ab ((cb ca ba) cb (ac ab cb))))
   ENTERING atomcount (s = ((ac ab cb) ac (ba bc ac)))
    ENTERING atomcount (s = (ac ab cb))
     ENTERING atomcount (s = ac)
     EXITING  atomcount (value: 1)
     ENTERING atomcount (s = (ab cb))
      ENTERING atomcount (s = ab)
      EXITING  atomcount (value: 1)
      ENTERING atomcount (s = (cb))
       ENTERING atomcount (s = cb)
       EXITING  atomcount (value: 1)
       ENTERING atomcount (s = nil)
       EXITING  atomcount (value: 0)
      EXITING atomcount (value: 1)
     EXITING atomcount (value: 2)
    EXITING  atomcount (value: 3)
    ENTERING atomcount (s = (ac (ba bc ac)))
     ENTERING atomcount (s = ac)
     EXITING  atomcount (value: 1)
     ENTERING atomcount (s = ((ba bc ac)))
      ENTERING atomcount (s = (ba bc ac))
       ENTERING atomcount (s = ba)
       EXITING  atomcount (value: 1)
       ENTERING atomcount (s = (bc ac))
        ENTERING atomcount (s = bc)
        EXITING  atomcount (value: 1)
        ENTERING atomcount (s = (ac))
         ENTERING atomcount (s = ac)
         EXITING  atomcount (value: 1)
         ENTERING atomcount (s = nil)
         EXITING  atomcount (value: 0)
        EXITING atomcount (value: 1)
       EXITING atomcount (value: 2)
      EXITING  atomcount (value: 3)
      ENTERING atomcount (s = nil)
      EXITING  atomcount (value: 0)
     EXITING atomcount (value: 3)
    EXITING atomcount (value: 4)
   EXITING  atomcount (value: 7)
   ENTERING atomcount
     (s = (ab ((cb ca ba) cb (ac ab cb))))
    ENTERING atomcount (s = ab)
    EXITING  atomcount (value: 1)
    ENTERING atomcount
      (s = (((cb ca ba) cb (ac ab cb))))
     ENTERING atomcount (s = ((cb ca ba) cb (ac ab cb)))
      ENTERING atomcount (s = (cb ca ba))
       ENTERING atomcount (s = cb)
       EXITING  atomcount (value: 1)
       ENTERING atomcount (s = (ca ba))
        ENTERING atomcount (s = ca)
        EXITING  atomcount (value: 1)
        ENTERING atomcount (s = (ba))
         ENTERING atomcount (s = ba)
         EXITING  atomcount (value: 1)
         ENTERING atomcount (s = nil)
         EXITING  atomcount (value: 0)
        EXITING atomcount (value: 1)
       EXITING atomcount (value: 2)
      EXITING  atomcount (value: 3)
      ENTERING atomcount (s = (cb (ac ab cb)))
       ENTERING atomcount (s = cb)
       EXITING  atomcount (value: 1)
       ENTERING atomcount (s = ((ac ab cb)))
        ENTERING atomcount (s = (ac ab cb))
         ENTERING atomcount (s = ac)
         EXITING  atomcount (value: 1)
         ENTERING atomcount (s = (ab cb))
          ENTERING atomcount (s = ab)
          EXITING  atomcount (value: 1)
          ENTERING atomcount (s = (cb))
           ENTERING atomcount (s = cb)
           EXITING  atomcount (value: 1)
           ENTERING atomcount (s = nil)
           EXITING  atomcount (value: 0)
          EXITING atomcount (value: 1)
         EXITING atomcount (value: 2)
        EXITING  atomcount (value: 3)
        ENTERING atomcount (s = nil)
        EXITING  atomcount (value: 0)
       EXITING atomcount (value: 3)
      EXITING atomcount (value: 4)
     EXITING  atomcount (value: 7)
     ENTERING atomcount (s = nil)
     EXITING  atomcount (value: 0)
    EXITING atomcount (value: 7)
   EXITING atomcount (value: 8)
  EXITING atomcount (value: 15)
15
—>
```

The Lisp function "atomcount" is traced as it is running on the list called "brahma"

Before a pilot is cleared to fly a new type of aircraft on routine passenger flights, he must undergo an exhaustive training course.

He must become so familiar with the controls and flight characteristics of the aeroplane that he could almost fly it in his sleep. And meet any emergency with the calm and confidence that is born of complete familiarity.

The cost of a single training flight, in terms of scarce resources, would appal a conservationist. As would the increased congestion and pollution in the vicinity of our already overcrowded airports.

A 747 uses over 3,000 gallons of fuel an hour at normal cruising speed. That's more than an average British motorist consumes in ten years of normal driving.

Fortunately for that motorist, Rediffusion has developed a generation of flight simulators so sophisticated that a pilot may learn to fly new aeroplanes without so much as leaving the ground.

Simulation, then, is one of the more dramatically rewarding applications of Rediffusion technology.

But we also experience a glow of corporate pride when passengers look up to check their flights on a Rediffusion Flight Information System.

Or when a ship's captain uses a Rediffusion navigation system to pinpoint his position with an accuracy that is almost uncanny.

Or, indeed, whenever we feel we have made a real contribution to the safety, comfort or convenience of people anywhere.

REDIFFUSION

THIS MAN HAS JUST FLOWN A BOEING 747 TO NEW YORK WITHOUT USING A DROP OF FUEL.

IF YOU WOULD LIKE TO KNOW MORE ABOUT US WRITE FOR A BROCHURE TO: REDIFFUSION LTD., CARLTON HOUSE, LOWER REGENT ST., LONDON SW1Y 4LS.

simple descent into a smooth-walled canyon and then a simple climb back up the other side; this recursion looks like a descent into a craggier canyon, where on your way up and down each wall you encounter various "subcanyons" that you must treat in the same way—and who knows how many levels of such structure you will be called on to deal with in your exploration? The shape described by a structure that goes on indefinitely like that has been named a "fractal" by Benoit Mandelbrot.

Notice in this recursion that we have more than one type of embryonic case (the "null" case and the "atom" case) and more than one way of descending toward the embryonic case (by way of both car and cdr). Thus our Big Questions can be revised a bit further:

1*a*. Is there just one embryonic case, or are there several cases, or is there even an infinite class of them?

1*b*. How can you know when you have reached an embryonic case?

1*c*. What are the answers to the various embryonic cases?

2*a*. From a typical case is there exactly one way to step toward an embryonic case, or are there various possibilities?

2*b*. From a typical case how do you determine which of the various routes toward an embryonic case to take?

2*c*. How do you build this case's answer out of the "magically given" answers to one or more simpler cases?

One of the most elegant recursions I know of originates with the famous disk-moving puzzle known variously as "Lucas's Tower," the "Tower of Hanoi" and the "Tower of Brahma." Apparently it was originated by the French mathematician Édouard Lucas in the 19th century. The legend that is popularly attached to the puzzle goes like this:

In the great temple of Brahma in Benares, on a brass plate under the dome that marks the center of the world, there are 64 disks of pure gold that the priests carry one at a time between three diamond needles according to Brahma's immutable law: No disk may be placed on a smaller disk. In the beginning of the world all 64 disks formed the Tower of Brahma on one needle. Now, however, the process of transfer of the tower from one needle to another is in midcourse. When the last disk is finally in place, once again forming the Tower of Brahma but on a different needle, then will come the end of the world and all will turn to dust.

A picture of the puzzle is shown below; I have labeled the three needles "a", "b" and "c".

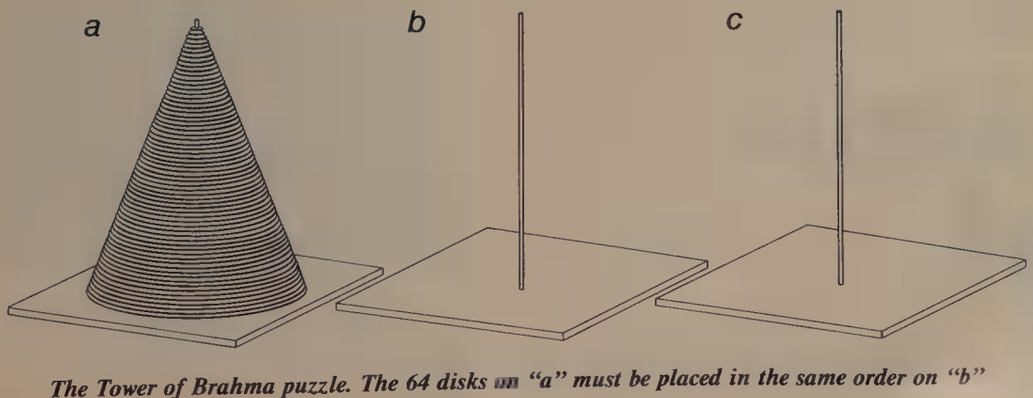
If you work at it, you certainly can discover the systematic method the priests must follow to get the disks from needle "a" to needle "b". For only three disks, for instance, it is not difficult to write down the order in which the moves go:

ab ac bc ab ca cb ab

Here the Lisp atom "ab" represents a jump from needle "a" to needle "b". There is a structure to what is going on, however, that is not revealed by a mere listing of such atoms. It is better revealed if one groups the atoms as follows:

ab ac bc ab ca cb ab

The first group of three accomplishes a transfer of a 2-tower from needle "a" to needle "c", thereby freeing up the largest disk. Then the middle move, "ab", picks up that big, heavy disk and carries it over from needle "a" to needle "b". The final group of three is much like the initial group in that it transfers the 2-tower back from needle "c" to needle "b". Thus the solution to moving three disks depends on being able to move two. Similarly, the solution to moving 64 disks depends on being able to move 63. Enough said? Now try to write a Lisp function that will give you a solution to the Tower of Brahma for n disks. (You may prefer to label the three needles with digits rather than letters, so that moves are represented by two-digit numbers such as 12.) I shall present the solution next month—unless, of course, the dedicated priests, working by day and by night to bring about the end of the world, should chance to reach their cherished goal before then.



The Tower of Brahma puzzle. The 64 disks on "a" must be placed in the same order on "b"

An IBM 3081: subatomic physics at Stanford

Finding the pieces in one billionth of a second.

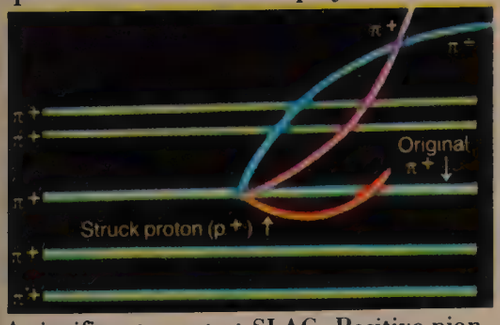
Trying to see how a watch works by slamming two together and examining the pieces is a popular analogy with physicists. It explains how they examine the innermost secrets of subatomic particles.

The watches represent matter and antimatter—electrons and positrons that whirl in opposite directions through a 1.3-mile ring at the Stanford Linear Accelerator Center (SLAC). Approaching the speed of light, they collide and generate bursts of subatomic shards that last for only a tiny sliver of time.

An IBM 3081 processor at SLAC has proved a smashing success in finding meaning in millions of bits

of data from these experiments, which may run for weeks or months.

As Charles Dickens, director of SLAC computing services, says, "Our 3081 moves data in and out of its central processor fast enough to provide the scores of physicists with immediate access to experimental results."



A significant event at SLAC. Positive pion (π^+) strikes proton (p^+) creating two new pions, one negative and one positive. Four other pions pass through with no contact.

They can think about the problem, and not worry about the computer, with microcode-assisted VM/CMS. With this flexible IBM software and the 3081's reliability, physicists can, with confidence, steal a look for a significant event during the experiment.

Dickens explains, "Before, people might say, 'If I'd only known what was happening, I'd have done things differently.' Now they see the results as they happen."

The 3081 supports 450 terminals for interactive computing under VM/CMS. It performs well in online environments because a new dyadic design allows its two separate processors to work concurrently on a wide range of applications. Its high reliability is due, in part, to an advanced thermal conduction module (TCM).

IBM offers engineering and scientific users extensive support: consultants, education and access to professionals. Tap into our many years of experience.

For an informative brochure, write Dr. Jack W. Hugus, IBM Engineering and Scientific Marketing, 1133 Westchester Avenue, White Plains, NY 10604.

IBM®

THE CAR:

New Renault Alliance. We set out to build the best small car in the world.

Affordable European technology. Our goal was an alliance of technology and affordability. Our achievement is Alliance, a fine European sedan–built in America–for the price of a four-passenger econobox.

Comfort for five. Alliance has eight inches more rear seat hip room than Escort. To comfortably ride five, not four. And with the front seats on pedestals, rear seat passengers can slide their feet forward and enjoy the ride.

Inspired performance. Front-wheel drive. Fully independent suspension. Power front disc brakes. Rack and pinion steering. Electronic ignition. All are standard.

Alliance's 1.4 litre engine is electronically fuel injected. Small wonder, then, that Alliance produces such outstanding mileage.**

52 EST HWY | 37 EPA EST MPG**

Alliance 2-door

$5595*

Built in America.

Now 11.9%
Annual Percentage Rate

Qualified buyers can now finance any new Renault, Jeep or AMC model at this new low rate. Jeep Wagoneer excluded. Applies to vehicles delivered through Mar. 31 or ordered by Feb. 28. Void where prohibited. Dealer contribution may affect price.

RENAULT
American Motors

THE AWARDS:

Motor Trend experts name Alliance 1983 Car of the Year.

Finalists were judged for styling and design, quality, comfort, ride and drive, handling, fuel economy, performance and dollar value.

"The Renault Alliance is the best blend of innovation, economy, and fun-to-drive we have seen in almost a decade. Moreover, it represents a uniquely successful blend of outstanding European engineering and American manufacturing know-how."
–Tony Swan, Editor
Motor Trend Magazine

And the European version of Alliance was named Europe's 1982 Car of the Year by 52 journalists from 16 countries.

Alliance DL 4-door

BUCKLE UP SAFETY IS A SNAP

*Manufacturer's suggested retail price for the Alliance 2-door. Price does not include tax, license, destination charges and other optional or regionally required equipment. **Compare 1983 EPA estimates for the Alliance 2-door with estimated MPG for other cars. Your actual mileage depends on speed, trip length and weather. Actual highway mileage and CA figures will probably be lower.

BOOKS

Eskimos in a modern milieu, bats, Gothic cathedrals, fractals, the heyday of steam

by Philip Morrison

THE LAST KINGS OF THULE, by Jean Malaurie. Translated from the French by Adrienne Foulke. E. P. Dutton, Inc. ($25.75). The Greenlander Knud Rasmussen, poet, ethnographer, explorer, the man "whose laughter went before him," holds sway over this book. His fancy named the place, his example led the author, his brooding face looks out of the frontispiece, and the small white house where once he lived still stands on the foreshore at Thule Air Force Base, before the huge phased-array radar and the big runways. But the kings of Thule, the polar Eskimos in archetype, the hunting people of that oasis along the northwestern corner of icy Greenland, have left forever. What they were like, how they came to their reign, how they live now and what the future may hold for them form the learned, anecdotal and heated chapters of this book of loving witness.

A remarkable history slowly emerges from behind the details of daily life and the dramatic personal encounters that crowd these pages. The place is a kind of second Bering. The complex conformation of the ocean channels has meant rich sea life and delayed freezing for so high a latitude. There is no easier place to cross from Canada to Greenland. The hunters found their way to the island continent at least as early as the 12th century. They knew land and sea, they took seal and walrus from kayak, snared the great flocks of guillemots, fished for fat salmon. The coasts are narrow in front of the ice cliffs; the people were only a few hundred in all the region, a couple of hundred miles of infolding shores. Then the climate changed. Between 1600 and 1800 a little ice age arrived, and the sea mammals became few. The birds were the mainstay, and the people suffered. In 1818 outsiders found them. Sir John Ross described them: without kayaks, without knowledge of how to take salmon or use bow and arrow, the old skills lost. Ross showed them wood for the first time. They had iron, some from meteorites lying on the coast, a little from Viking trade. From then on whaling ships came by occasionally, with wood and iron.

In about 1860 the sorcerer Qillarsuaq, who lived among the Canadian islands more than 500 miles to the south, was "seized by the prescient certainty" that much farther north were living hunters still unknown. He persuaded his own people to join him for an indescribably difficult journey into the unknown. Had his "white hair not been haloed by a nimbus of fire every night" who would have stayed the course for seven seasons? In the end they found lots of men and women and their sledges. The Canadian Eskimos settled in the new place, reviving the arts of kayak and bow.

Now came the polar searchers, Peary and Cook. Peary came every five years or so, engaged half of the hunters to staff his journeys, took back the great meteorites to New York, traded for walrus ivory and fox furs, on terms of trade most unfavorable to the hunters. After 20 years Peary was a hero to the Inuit, even though he had never dealt fairly with them. In his decisive way, however, he had brought them rifles and ammunition. In 1951 Malaurie met the two sons whom Peary and his long-time companion, the black seaman Matthew Henson, had left behind. There are persuasive photographs of these two hunters, men ignored in the U.S.

After Peary's somewhat uncertain claim on the Pole he returned no more. Now it was Rasmussen's time. He established the trading post at Thule—so he styled the old Inuit village of Uummannaq, the seal's heart—to supply the hunters, taking pay in the beautiful blue-fox fur they brought in proudly. The Danish administration did not want to extend its trading and service network that far north. Until Rasmussen died in 1933 he was in effect the government of the people, and the proceeds of their trade supported his seven Thule expeditions that laid the foundations of Eskimo history. Some 300 polar Eskimos financed the finest program of studies of their own past, over a front of 9,000 miles. The hunters revere him still; he was one of them, lofty of purpose, respectful of dignity, entirely trustworthy.

It was in 1950 that Malaurie came to the people of Thule, straight from the Sahara, where he had been mapping with Tuareg guides. He was a geomorphologist with a strong desire to know the people of the ends of the earth. His project was a careful aneroid survey of the beaches of the region, as a contribution to the understanding of the relaxation of glacial stresses. The Inuit were sophisticated enough by 1950, although they hunted still for a living. They had never, however, heard French. They liked the tongue, the fables of La Fontaine and the robust and competent young scientist with such a strong sense of character. He published his survey maps, which had cost a long, hard journey with a party of hunters and their families, but geology was now behind him. Since then he has been a student, companion and filmmaker for Eskimos around the circumpolar world.

As Malaurie came to Thule on his homeward way he encountered an event that cut across the history of the region with the decisiveness of a slip along some great fault. In the summer of 1951 the lonely bay at Thule became host to an unprecedented flotilla of ships; the shore was forested with construction cranes; 20,000 men and their machines arrived to work for the U.S. Strategic Air Command. By 1954 the government of Denmark had built the hunters a new village to the north, where they live today. The SAC base holds 3,000 airmen still. The blue foxes live off the generous Air Force dump. The guards shoot them as vermin now, the pelts of no interest.

Elsewhere the foxes are scarcer. And the kings of Thule? The hunters stay home; they feed on tea and sugar, with smokes and drinks in plenty. At the age of 50 those who were hunters in 1951 can no longer chew meat. Danish welfare has replaced the arduous hunt.

It need not be. Indeed, hope is present. By 1980 it looked like "Eskimo Gaullism," Inuit power. Men of Thule came to the all-Greenland meeting in the south proudly wearing the rare bearskin pants. The choice is not the isolated past of pride and terror, the kings vassal to the icy wind, infants sacrificed in dire need. Pay a fair price for the lovely furs. If the fox range must be extended, haul the hunters out by plane. They can return on their own, hunting as they go, the old skills augmented by the rifles, radios and compasses they can use. Support a strong emergency patrol system, the hunters out in the coldest times, maintaining a search-and-rescue network that might someday save the lives of air passengers downed along their great-circle route between two warmer climes. It may be that some form of husbandry for foxes and bears and seals can be developed, intermediate between the open hunt and the mink ranch. It will all cost the taxpayers of Copenhagen less than the welfare state, and it can allow those Inuit who do not treasure continuity to seek a newer way.

It might also preserve the kingship,

You can tell from the outside
which Scotch they serve on the inside.
Johnnie Walker
Black Label Scotch
OLD
12 YEAR OLD BLENDED SCOTCH WHISKY, 86.8 PROOF. BOTTLED IN SCOTLAND. IMPORTED BY

the royal view that "we are the strongest" or the phrase that "for an Eskimo it's easy." Like the kings of other latitudes, these nomads of Thule are not always admirable. They are often vain, mercurial, harsh. Their consorts, the woman tied to the hunter both by a strong sexuality and by the most urgent necessities of mutual survival, were granted dignity and status, but they lived too hard a life, worn out long before their men. We cannot preserve all the past, but this human past is symbolic for us all; did not our species embark on its strange, swift adventure once the edge of the ice receded?

ECOLOGY OF BATS, edited by Thomas H. Kunz. Plenum Press ($49.50). Although they are furry, breast-fed kin of ours, the bats have a bad press, whereas their feathered reptilian predecessors in the air catch all eyes. The reason is in the eye itself; bats are in a way the nocturnal equivalent of birds. They are successful aeronauts everywhere, diverse and abundant except at sea and toward the poles. Their world, however, is dark and colorless, so that we diurnal primates can watch it little.

This volume is a collection of 10 expert reviews of what bat watchers have lately come to know. The past 20 years have placed this order of nearly 1,000 species closer to its proper position in our understanding: it is the second order among mammals. (There are still more rodents than bats.) Bat studies are not easy; night work is the rule, with snooperscopes or telemetry receivers, with sonar translators or deep in caves filled with guano. In the laboratory too bats are taxing subjects. The book, whose level is that of the research survey in a set of well-written examples, serves two purposes for the general reader: it gives a clear account of its topic (if one heavy in taxonomic jargon) and it presents an exhibit of the detailed, mathematicized, ambitious (and still incomplete) state of field biology today.

True to its title, the collection tends to look away from the form and function of single bats. There is little on flight aerodynamics (perhaps too little) and almost nothing on the sonar system: the fierce mouths for pulsing shrieks and the labyrinthine directional ears. Roosts and fleas, reproduction, growth and survivorship, energetics and size relations, feeding strategies for both insects and fruit—these are the topics of comparison and conjecture. A motif that recurs, rather wistfully, is the desire to determine ecological niches from the form and ability of species. It is hard to see how it is that so many distinct species with such distinct pulses hunt so well the same mixed bag of bugs in flight. There is a lot going on we do not know.

Most of the larger Old World plant-visiting bats lack echolocation. The biggest have a wingspread close to six feet. They roost externally, often in the very trees where they feed in the warm tropical nights. Mousy Temperate Zone bats, gregarious in their caves by the 10 million, or widespread in the belfries, are insectivores, demon echo-locators of insects by dawn and dusk. Their enclosed roosts call for fancy navigational skills without light but reward them by keeping out the freezing chill of winter. Birds that in warm weather live on the insect hordes generally migrate a long way toward the Tropics once cold weather ends the buzzing life for the season. The bat counterparts migrate in time rather than space; they take the winter off not in Florida but in hibernation, a complex and well-studied state, energy-economic because the animal's temperature approaches the ambient one. This strategic choice is not yet understood.

In flying at night bats avoid the fiercest birds of prey. It has lately been recognized that owls, the silent night hunters, take an important toll of bats. The deeds have been witnessed from blinds with the infrared image amplifier, and the search of owl pellets for bat bones has confirmed the indictment. In the rich study of bat and light rhythms reported here one striking graph shows the measured activity of two bat species and two owl species as a function of light levels. There is overlap, because bats need enough night hours of work to make a living. But the peaks are sharply displaced: bats prefer to work at the light level of the dark of the moon, owls at that of the full moon. During a recent lunar eclipse the activity of bats showed an unexpected small-hours peak.

Owls do not hunt bats by listening, as they hunt mice in response to the scurry in dry leaves. In order to capture a fast-moving, darting bat the owl must track it at longer distance and strike swiftly. Deep darkness is the shield of the echolocating bat, even though the dark cannot hide the bat's insect targets from the sonic pulse or confuse its quick flight into the twisted crevices of home. In this connection the bat mother in the metropolis of the bat cave does not seek out her own infant. She nurses the first hungry kit she can find; her extended family can number in the millions.

Bats seem to live long for their size, both actuarially and physiologically. No fossils disclose any prebat stage before bathood. Perhaps the bats originated with a tiny tropical omnivore of the trees, first giving rise to a small flying pursuer of insects, then radiating into larger forms able to live on fruit, nectar or meat. The animals later developed a complex system of thermoregulation, and they could reproduce somewhat independently of external conditions. The hibernators of our time are to be seen as late-arising specialists of the seasonal world outside the Tropics. The cave is fully home now to the little bats, as much later it became a refuge and a home to our own lineage.

EXPERIMENTS IN GOTHIC STRUCTURE, by Robert Mark. The MIT Press ($15). The tall pages of this not very large book are fitting for the beautiful tall forms that stand on many of them: elegant sections of cathedral naves in sharp line, photographs of the entire masonry fabric of those great buildings, sections of plastic models mapped in the rainbow colors of stress interferometry. The argument too is elegant; this Princeton engineer on the site and in the model laboratory has probed for understanding behind the swift progress of the master builders of medieval Europe. Those designers and the teams they led, not serfs but "free—nay—unionized" masons and carpenters, learned in two centuries to capture ever more soaring space in their common stone.

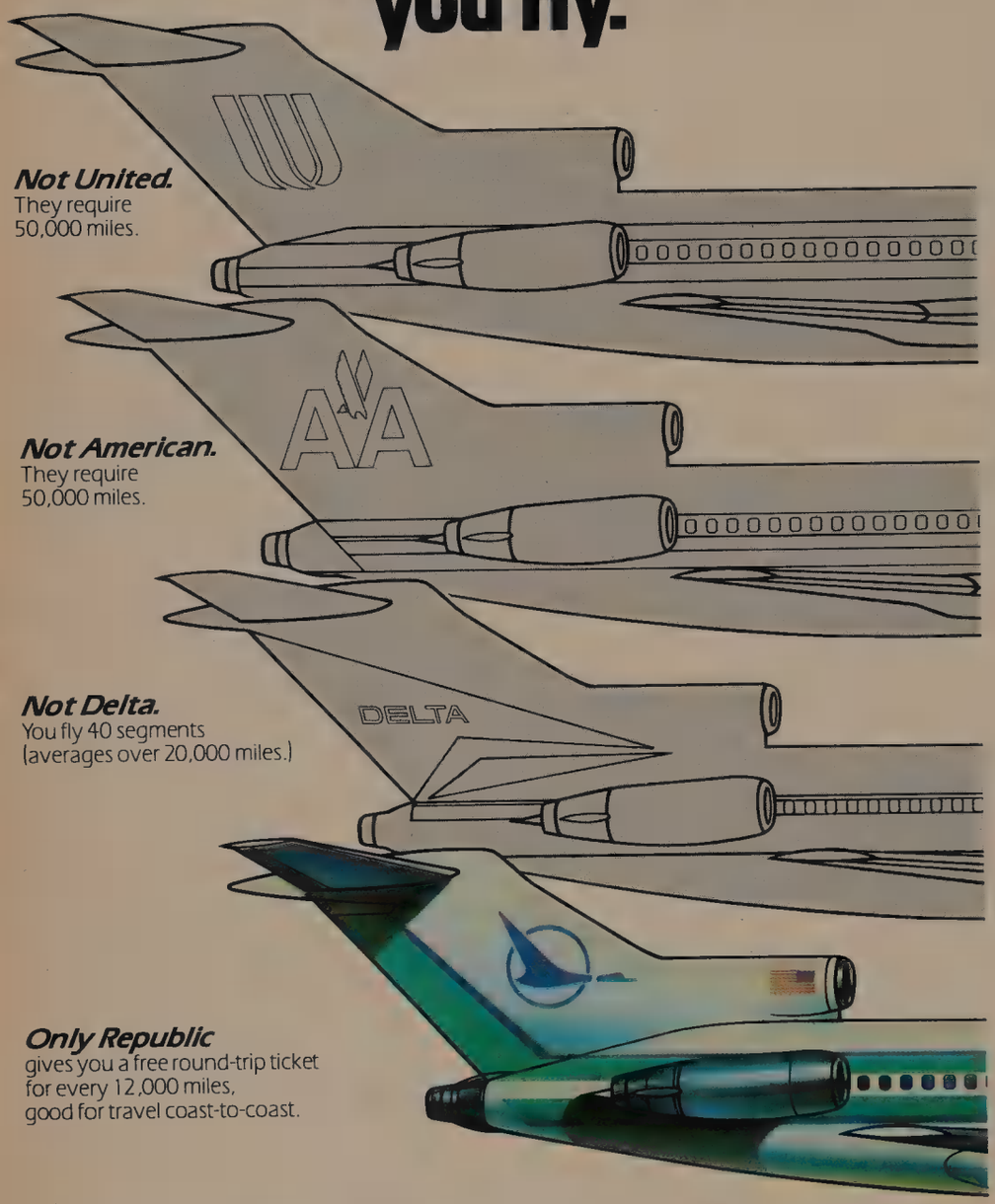
The issue is not a new one. For a century the architect-restorers (Viollet-le-Duc is the most renowned of them) and the historians have joined over it. Structure, however, is not simple. Only recently have we been able to examine in some detail examples geometrically complicated enough to be relevant. Insights alone give equivocal answers to many of the questions the structures raise; Mark's models allow judgments that archaeology and argument cannot by themselves decide.

Mortar-jointed masonry gave rise to progress: it made possible the great towers and aisles. Even the highest belfry does not out-top the solid Pyramid of Cheops; stone on this scale does not fail under compressive load alone. But simple stability, each block in balance on the block below, does not change with scale. A model of toy blocks could serve for the lofty cathedral of Reims; indeed, such models may have been used. More important, the builder's experience would stand him well as his ambitions and structures grew. Masonry supports little tension; if the forces impose tensile stress, the masonry will soon fail visibly. Localized areas of tension cracks give a clear sign of incipient instability.

So armed, the builders could build with understanding, although surely they had no mechanical theories. The flying buttresses in building after building show their virtues in model tests, as for a long time the analysts have expected. Nobody "of sound mind" could take away the fliers from Reims, just as no one would now remove them from Nervi's ferroconcrete shells in Rome.

As the churches towered higher, the lateral wind loads became more important. The models disclose the entry of intolerable although localized tensile stress under realistic if unusual winds. From this we can infer the old builder's care in watching for such signs, and his

Nobody
but Republic gives you free travel for every 12,000 miles you fly.



Not United.
They require
50,000 miles.

Not American.
They require
50,000 miles.

Not Delta.
You fly 40 segments
(averages over 20,000 miles.)

Only Republic
gives you a free round-trip ticket
for every 12,000 miles,
good for travel coast-to-coast.

**The new Republic Frequent Flyer™ Program:
enroll now and begin earning your free round-trip travel.**

Name _____ Company _____

Phone: Home ( ) _____ Business ( ) _____

Do you prefer to receive your Frequent Flyer mail at: ☐ Home ☐ Business

Address _____

City _____ State _____ Zip _____

Mail to Republic Airlines Frequent Flyer Program, Dept. SA , P.O. Box 20452, Atlanta, GA 30325.
We'll forward your personal account number and full information.

NOBODY SERVES OUR REPUBLIC LIKE REPUBLIC™ AIRLINES

modification of the design to cure it. The ornate pinnacles that weigh down the high piers are often—not always—effective measures against such failure. Moreover, it is clear that a cathedral is no fixed structure; constant repair today is evidence of the passing seasons in these economical and open designs. Beauvais, at 50 meters the highest of all towers, collapsed twice, once before 1300 and again in about 1570 (when it had been audaciously rebuilt with the tower 150 meters tall). The models suggest (the inference cannot be final because the structure is not well recorded) that the wind loads induced cracking within a poorly visible set of intermediate pier buttresses.

The art continued; Palma Cathedral is much more spacious than Reims. Its nave is more than twice as high; by one measure it is twice as slender. Heavy external pier buttresses, artfully arranged to house side chapels, have opened the interior safely without loss of the visual harmonies. Vaulting is still more difficult to analyze without models; even the ribbed barrels of the St. Louis airport, let alone the wonderful ceilings of the French High Gothic, do not disclose their structural necessities in visible form.

When Professor Mark was a college engineering student, he incautiously carried a volume on Picasso to help him wait out an early interview. A sympathetic engineer observed: "Young man, never bring a book like *that* when you look for a job in engineering." This book shows his independence of mind. He has not, however, forgotten the broader demands of unity: both the beauty of the volume and the careful explanatory apparatus—not one equation but plenty of serious mechanics—fit it alike for humanly curious readers with engineering savvy and for those who deal freely only with word and form but remain open to the support of quantitative reason.

THE FRACTAL GEOMETRY OF NATURE, by Benoit B. Mandelbrot. W. H. Freeman and Company ($32.50). In 1975 there appeared a small but dense and delightful volume in French around the neologism "fractal," the invention of this gifted, playful and tirelessly imaginative author. That work, "macédoine de livre," was reviewed in these columns with the enthusiasm it deserved. It was promptly enough served up in English, the key word still visible, the book fuller, much more visual, extended in scope and form. Here it is once again, the new edition straightforward of title, bearing a set of color plates that exhibit the virtuosity of today's programmers, the mixture as fragrant and fresh as before.

The new form, perhaps increased a third in size, is ready for lucky readers who have not yet found their way to fractals, no less than for those who sim-

ply want more. Meanwhile the topic has become all but trendy. Strange attractors and renormalization groups on a lattice, topics that blaze in the physics colloquiums these days, are here, full of evocative content albeit incomplete. The subject matter of the older editions is not neglected: the simplest idea behind the concept leads still to the understanding that the coastline of Maine has a length dependent on the inquirer. A ruler laid on the map from Portland to Calais gives a minimum answer; a line drawn through the indentations of bays and headlands on a highway map gives a longer one; the ruler put down centimeter by centimeter at real tidewater gives a much longer one, including every cobble and pebble.

The point is made quantitatively and is related to the notion of dimension. Such an exquisitely tortuous curve falls between the dimensionality of a line and that of a plane, reasonably enough. Hence the idea and the word "fractal." Of course, chance must enter, although rich deterministic examples are found in the dizzying idealized constructions of the long-known theory of real variables. Let the matter rest there; those who wish can find their way to the subject easily.

The book is no text, no set of definitions, no series of applications. It is a flowing source, as once the wonderful *On Growth and Form* by D'Arcy Wentworth Thompson came to us all. It too belongs on the small shelf of books that disclose the forms of nature. Thompson, classical in approach, centers on the simple forms hidden in the complex web of life. Mandelbrot, resting squarely on the mathematical moderns—Cantor, Koch, Hausdorff, Paul Lévy—opens instead a path to the simplicity of plenitude. His heroic motifs are not spheres or spirals but intricate ideals of islands and reefs, stars and curdled milk, turbulent spray and pink noise, the abstract forms of nature's generative profligacy. Naturally enough, the book itself is less lucid, but it is unifying, recursively fascinating, agreeably personal without posturing and buoyant with scientific hope.

THE POWER OF STEAM: AN ILLUSTRATED HISTORY OF THE WORLD'S STEAM AGE, by Asa Briggs. University of Chicago Press ($22.50). Frontispiece and jacket bear the title of the volume in an artful style: the legend is a cast brass nameplate, evocative of the Steam Age less only than the hiss, the throb and the scent of oily rags. From a photograph of a Newcomen engine in place (1855), past the Exhibition of 1851, the wire mill of Sharp and Brown, a Victorian smokestack topped with Norman crenelations, the *Great Eastern* in New York to turbines of our day the illustrations present an overview both canonical and fresh. Sooty miners and spruce bellhops,

HOW TO GAIN ACCESS TO THE COMPLETE SCIENTIFIC MEMORANDA OF MIT AND STANFORD'S ARTIFICIAL INTELLIGENCE LABORATORIES.

Not just the findings but the complete memos

The development of Artificial Intelligence took place at two great AI Laboratories. At MIT under the direction of Professor Marvin Minsky and at Stanford under the guidance of Professor John McCarthy. The complete, unabridged sets of reports, memos, idea exchanges and research findings at these two great laboratories is now available for the first time ever as a collection in two new sets of microfiche. The MIT set covers 1958-1979. The Stanford set covers 1963-1982. There's a special introduction to the MIT set by Professor Minsky. And over 10,000 pages of memoranda are represented.

Over 800 reports organized, indexed and catalogued.

The two breakthrough collections are carefully designed for research use. A fully annotated, indexed catalogue with abstracts covers such topics as seminal work in robots, LISP, expert problem solving, understanding language, vision, theories, systems and games. Retrieving reports or specific information has been greatly simplified by a team of top editors and research professionals. This important collection is a simple and easy to use reference.

Essential for AI and Robotics libraries and laboratories.

As a research tool, as a reference tool or simply as fascinating reading, these two first time collections will undoubtedly be in great demand. They are a must for the shelves of any institution, university or corporation that considers itself an entity in the world of Artificial Intelligence or Robotics. And they are clearly destined to be scientific classics to be used by engineers, experimental psychologists, linguists, computer scientists as well as AI professionals.

MIT and Stanford: The basis for all future developments in AI.

MIT, Stanford, Professor Minsky and Professor John McCarthy took a long but rewarding voyage into the exploration of thinking. On the way, they discovered more than they ever dreamed of. Now, you can discover the institutions and personalities that had the vision to see what the study of Artificial Intelligence would bring to the 80's. Share those great, scientific moments in two exciting microfiche collections. The MIT collection (1958-1979) sells for $2,450. The Stanford collection (1963-1982) sells for $2,750. For orders or more information, clip the coupon below. Or better yet, call us Collect at 212-838-7200. Ask for Department 101A.

Comtex Scientific Corporation
Department 101A

850 Third Avenue
New York, N.Y. 10022

☐ Please send _____ set(s) of the AI/MIT Memoranda (1958-1979) @ $2,450 per set.

☐ Please send _____ set(s) of the AI/Stanford Memoranda (1963-1982) @ $2,750 per set.

☐ Please send me more information about the Memoranda.

☐ I'd like to be kept up to date with Comtex publications. Please put me on your mailing list.

Name_____

Organization_____ Title_____

Address_____

City_____ State_____ Zip_____

Please feel free to call us Collect at 212-838-7200. Ask for Department 101A.

Close your eyes. Now have someone read this to you.

You are blind. A student. Facing four years of college.
With about thirty-two textbooks to read. Plus fifty supplemental texts.
How are you going to manage?

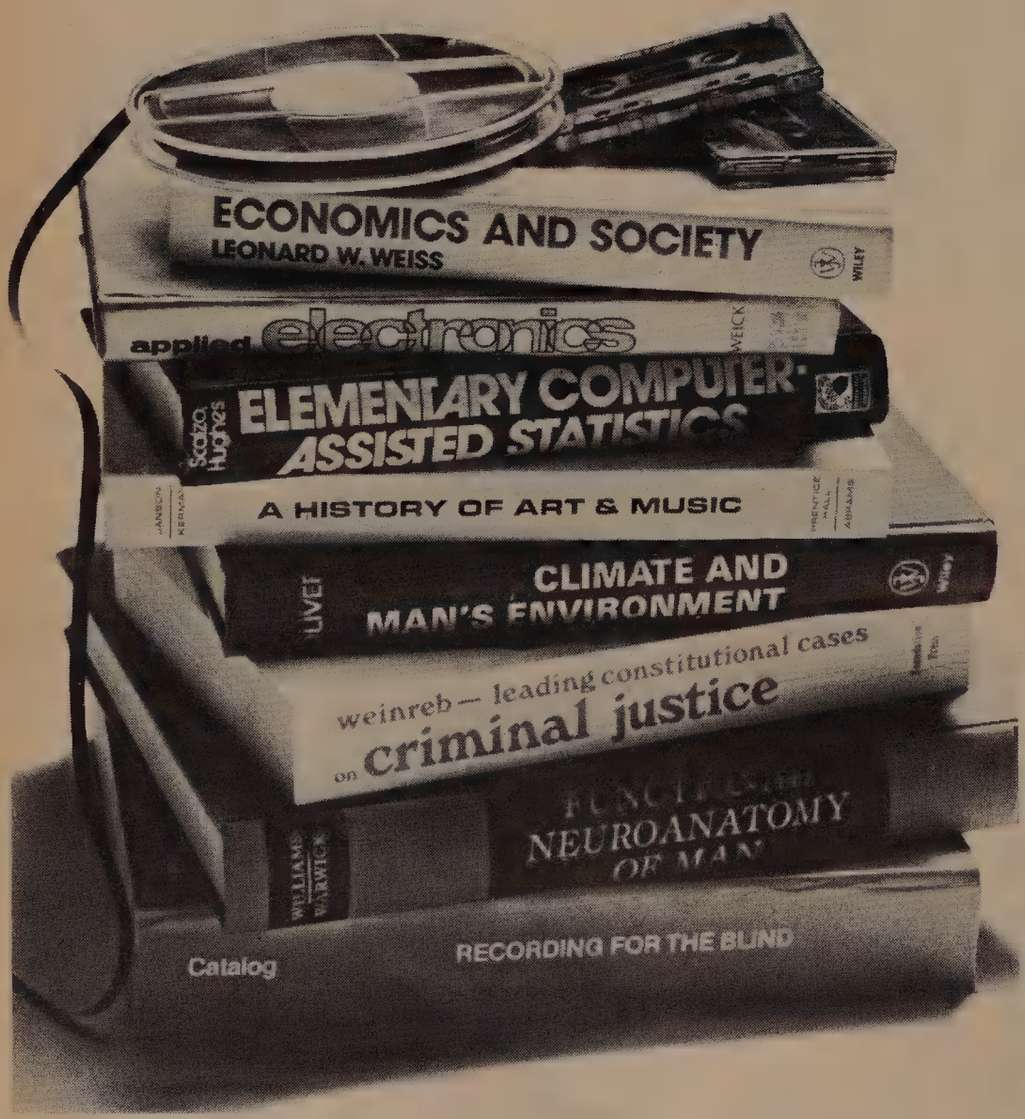
With Recording for the Blind. Since 1951, we've helped over 53,000 blind, perceptually and physically handicapped students get through school.
By sending them recordings of the books they need to read. Free.

Recording for the Blind is non-profit, and supported by volunteers and contributions from people like you who can imagine what it's like to be blind.

Your tax-deductible donation will help our students meet their educational goals. We'd all be grateful.

If you want to know more about us, write:

Recording for the Blind, Inc.,
an educational lifeline.
Station E, 215 East 58th Street., New York, New York 10022,
(212) 751-0860.



sheet music and racing paddle-wheelers, the view is a wide one.

The book seeks to "bridge the gulf" between the pictorial albums and the austere volumes of economic history. Lord Briggs is a distinguished cultural historian of the period; his text is supple and persuasive, although it is more a fragrant broth than a savory hot pot. The beginnings are particularly well put. Steam power, almost a century before James Watt, served to pump the big collieries dry; England was a smoky industrial country before the Industrial Revolution. Textile factories, particularly for the fulling of woolen cloth, were users of considerable power long before Watt, driven by water and even by horses. Watt rated his engines, after all, in horse power, a comparison quite familiar to his customers. Karl Marx said it well: "The steam engine itself... did not give rise to any industrial revolution. It was... the invention of machines that made a revolution in the form of steam engines necessary."

Steam power, both the stationary engines in every shop and mill and the locomotive engines by land and sea, gave a kind of unity to the rise of industry and to its diffusion worldwide. Lewis Mumford spoke of "carboniferous capitalism." For Briggs there arose rather a "gospel of steam," a vision of global progress through pistons. What happened was worldwide, but it was at once both richer and crueler than the poets and the reformers could have foreseen.

The book ends in our time of internal combustion and electronics. Locomotives and tractors are diesels these days; the reciprocating steam engine has become a piece of archaeology. The text does make the point that spinning turbines under high-pressure steam still dominate the big power plants, but it does not enough emphasize that the bulk of the electric power of our world is still the work of steam turning steel shafts. Indeed, the last chapter is frank nostalgia: vintage steam. In Britain there are some 200 sites where pistons under steam can be seen alive and at work today or in a specialized museum setting. About 30 sites are named in the U.S., and others are specified worldwide.

Asa Briggs is chancellor of the Open University. Although he has read Henry Adams, he seems to cavil at the late Lord Snow's plausible remark that the laws of thermodynamics might be held as material even a literary person ought to know. Laws narrowly "drawn up to explain the workings of steam engines" in fact hold universally, and cast the coldest of eyes on life, on death. The two cultures still walk two paths. This book was conceived and assembled by a clever British editorial firm. The absence of meaningful sources for the many fine illustrations is the only deficiency of this attractive and readable volume.

Jong Seh Lee came here to learn city planning. We think he's learning a whole lot more.

For Jong Seh Lee, the shortest path between his home in Seoul, Korea and his chosen career went through the University of Pennsylvania, in Philadelphia.

There, he's pursuing graduate studies in urban engineering with some of the world's foremost authorities on the subject.

For Jong Seh Lee, studying in this country was a long-cherished dream. And making it possible is an ITT International Fellowship.

Jong Seh Lee was one of 51 Fellows last year—half of them foreign graduates who came here to study, the other half Americans who went abroad.

Since this innovative fellowship program began nine years ago, over 500 young men and women have been helped along in their careers, which have ranged from teaching of history to the engineering of modern architecture.

To these gifted youngsters, ITT Fellowships offered a unique opportunity to learn more about their chosen fields.

But to us, their ITT Fellowships have always had still another purpose: to give some of the world's future leaders an opportunity to learn something about Americans.

And equally important, Americans about them.

The best ideas are the ideas that help people. ITT

© 1982 International Telephone and Telegraph Corporation, 320 Park Avenue, New York, N.Y. 10022.

Established 1845 SCIENTIFIC AMERICAN March 1983 Volume 248 Number 3

The Law of the Sea

In December 119 nations signed the Convention on the Law of the Sea. The U.S. was not among them, but development programs under the new law will nonetheless be proceeding

by Elisabeth Mann Borgese

On December 6 of last year the representatives of 119 nations, gathered at Montego Bay in Jamaica, signed the United Nations Convention on the Law of the Sea. The ceremony culminated 15 years of labor by the largest and longest of all international conferences. The Third United Nations Conference on the Law of the Sea had engaged as many as 3,000 delegates in 11 long sessions totaling 585 days. Beginning in the chaos and conflict of contending political and economic interests, working through draft after draft, the Conference succeeded in writing a "constitution for the oceans" in 320 articles (organized in 17 main parts) and nine technical annexes.

If the gist of this enormous work can be stated in one sentence, the Convention replaces the traditional laissez faire system of freedom of the seas with an emerging system of management. In place of the two-dimensional boundless sea the Convention deals with a finite three-dimensional resource, its depths of as much economic interest as its surface. It puts 40 percent of the ocean and its bottom adjacent to the coasts of the continents and islands under the management of the states in possession of those coasts. It reserves the other 60 percent of the surface and the water below for the traditional freedom of the seas, but it deeds the wealth of the ocean floor, 42 percent of the earth's surface, to the Common Heritage of Mankind. The Convention places that heritage under the management of an entirely novel International Seabed Authority, which has the capacity to generate income, the power of taxation and a kind of eminent domain over ocean-exploiting technology.

Out of opposition to the Seabed Authority, it is true, the U.S., the U.S.S.R. and 15 other major industrial nations withheld their signatures from the Convention. These nations did, however, sign the Final Act of the Conference. (Only one nation, Turkey, signed neither.) The nations that signed the Final Act but not the Convention are entitled to attend future meetings of the agencies created by the Convention as nonparticipating observers. The signing of the Convention gives it the force of law, subject to ratification by at least 60 nations. When the Convention has that ratification, it can be expected that the nonsigners will sign as full participants.

The Convention provides a comprehensive global framework for the protection of the marine environment, a new regime for marine scientific research and a comprehensive system for settling disputes. It ensures freedom of navigation and free passage through straits used for international navigation, invoking the new legal concept of "transit passage," a right that cannot be suspended under any circumstances. It updates and codifies traditional law. It defines rights as well as responsibilities for states and international organizations in establishing and enhancing a new order in the world's seas and oceans.

The making of the new Law of the Sea engages all the issues the world community has to face in our age: food and energy, the arms race, communications and the "information revolution," trade, commodity policy, resource management and conservation, regional integration and the movements in science and technology that underlie many of these. It is as though all the issues were flowing together in the world ocean and the ocean had now become our laboratory for the building of a new world order. That order, we may hope, will prove more rational, more humane and more responsive to the real needs of the world than the old order that is disintegrating in hunger and violence.

Until 1973, when the Third United Nations Conference on the Law of the Sea began, the law of the sea had always been made by one or two or half a dozen seafaring states. Through the centuries, or the millenniums, the old-world seas came under the shipping code of Hammurabi and later the Rhodian sea law. The Mediterranean, the cradle of Occidental civilization, became the *mare nostrum* of Rome. In 1493 Spain and Portugal divided between them the world ocean beyond the Strait of Gibraltar. For the maritime power of 17th-century Britain, John Selden propounded the theory of *mare clausum,* the closed sea "owned" by the coastal states that would have the right, like the Spanish and Portuguese in their day, to control and exact tolls from navigators of other nations. It was against this doctrine that the Dutch jurist Hugo Grotius advanced the concept of *mare liberum,* the free ocean open to navigators from all nations, each respecting that right for others. In historical reality the freedom of the seas could be enjoyed only by those who could enforce it; throughout modern history they have been a handful of countries of western Europe and North America.

Two dramatic developments in human affairs, concurrent and often conflicting, compelled the calling of the Third United Nations Conference on the Law of the Sea. The first is the maritime dimension of the scientific-industrial revolution that is painfully transforming industrial societies and profoundly affecting their relations with preindustrial societies. The discovery of the phenomenon of sea-floor spread-

ing has had the earthshaking, in the literal sense, consequence of installing plate tectonics as the central organizing idea in geology. That same discovery has far-reaching economic implications. Where the sea floor is rifting, at the axes of its spreading, metals selectively separated from the mantle of the earth well up to the surface. It is now understood that this process laid down the veins of nonferrous metals hitherto worked in the world's hard-rock mines. From the same closer acquaintance with the sea floor we know that it holds more oil and gas, in the continental shelves and in the benthic floor as well, than the traps in continental rock (much of it too, of course, once sea floor).

Meanwhile the rapid technological development of the fishing industry has undone the old notion that the ocean fishery is inexhaustible. The assumed infinity of the ocean has been, at the same time, overtaxed by pollution from urbanization, industrialization and rising population.

Thus the assumptions of the old law of the sea were eroded and the law began to crumble. The U.S. started the expansion of national claims to the ocean and its resources in 1945, with the Truman Proclamations on the Continental Shelf and the Extended Fisheries Zone. A scramble was on to carve up the world ocean, just as Africa had been carved up a century earlier.

While the Industrial Revolution was penetrating the ocean from the Northern Hemisphere the second great impulse that has shaped the new Law of the Sea was gathering force in the Southern Hemisphere. From the colonial empires assembled by the former maritime powers of Europe there came a procession of new sovereign nations. No more than 70 nations had participated in the First United Nations Conference on the Law of the Sea, called in 1958, or in the second one in 1960. By 1973, when the Third Conference convened, the membership of the UN had more than doubled. The majority in the General Assembly had passed to the poor nations of Asia, Africa and Latin America. They were determined to have their say in the deliberations of the Conference. Their call for a new international economic order was about to be sounded. The new Law of the Sea was to be part of it and even a model for it.

The Convention on the Law of the Sea clearly exhibits the shaping by the parallel yet often intersecting revolutions in technology and international relations. It was the technological revolution that pushed from the North the expansion of the national jurisdiction in ocean space. It was the revolution in international relations that pressed, from the South, for the establishment of new, strong international institutions. The two revolutions are equally in evidence in the massive record of the proceedings of the long conference. The Convention itself and the Conference record also reflect, however, the presence of a number of able, scholarly and statesmanlike personalities from both the North and the South: Arvid Pardo of the island state of Malta; Hamilton Shirley Amerasinghe of Sri Lanka, president of the Conference until his untimely death in 1981; Jens Evensen of Norway; Jorge Castañeda of Mexico; S. P. Jagota of India, and Louis Sohn of the U.S., to name only a few. The workings of their individual genius interacted with the collective pressures from the North and the South to steer the work of the Conference.

It was Pardo who proposed, in a historic address before the First Committee of the United Nations in 1967, that the wealth of the ocean floor should be declared the Common Heritage of Mankind. Neither the freedom of the sea nor the sovereignty then being claimed on the ocean by coastal states, he said, could ban the triple specter of pollution, of exhaustion of marine life and of international strife. Only international cooperation could preserve the peace and bounty of the oceans, and this might be furthered by the circumstance that vested interests were not yet as firmly entrenched in the oceans as they were on land.

A Committee on the Peaceful Uses of the Seabed was immediately appointed; a Declaration of Principles, incorporating Pardo's major thesis, was adopted in 1970, and the reservation of the seabed for peaceful purposes was incorporated in 1972 in the treaty that bans nuclear weapons from the seabed. In 1972 the seabed committee, still at work, called for the convening of the Third United Nations Conference on the Law of the Sea and began the preparation of its agenda.

The agenda was completed in 1973, when the work of the Conference on the Law of the Sea began. Organized in a General Committee, three main working committees and an intricate network of regional groups, interest groups, negotiating groups, plenaries, informal plenaries and a drafting committee working in six languages, it was the biggest and most complex conference in history.

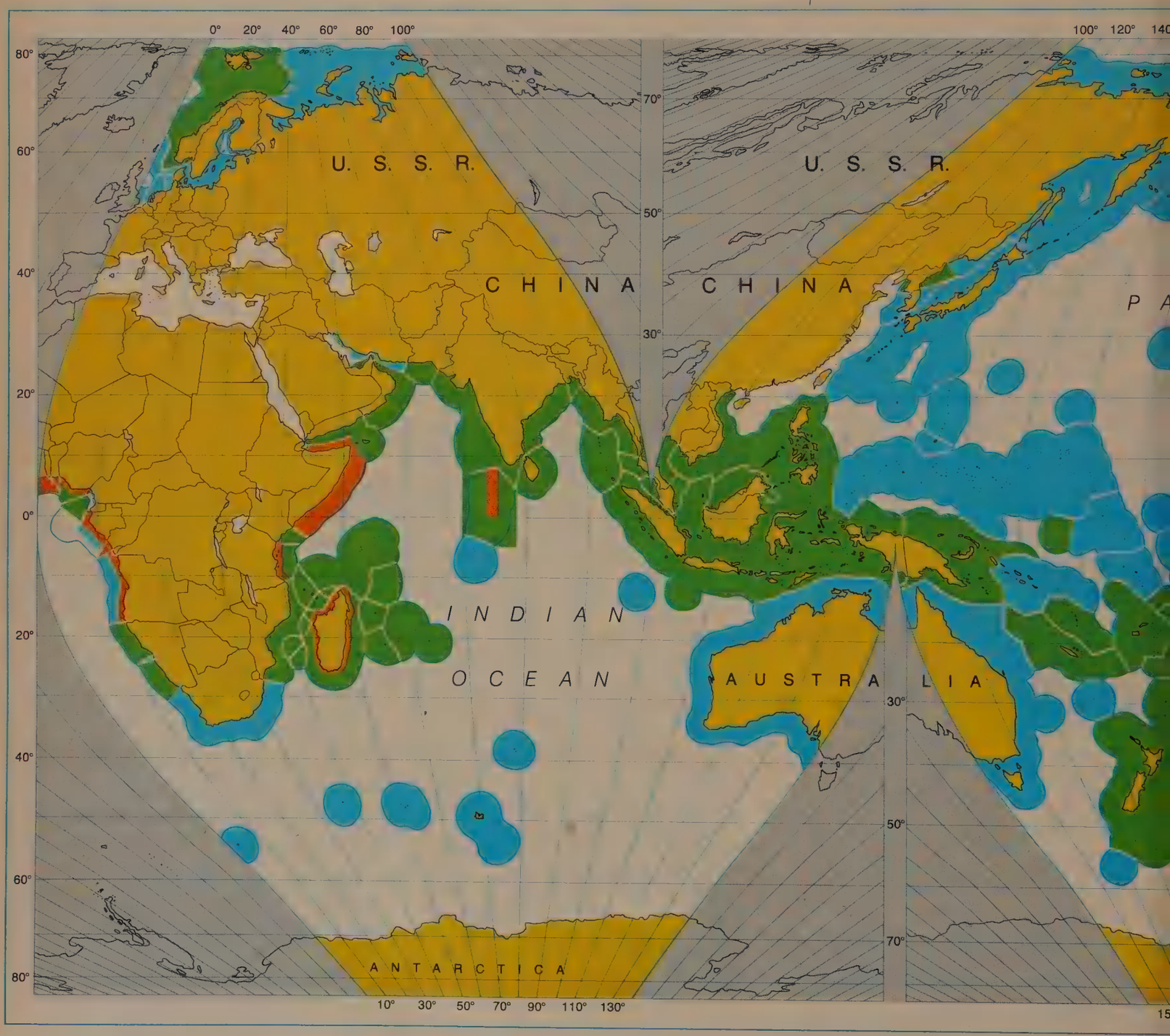
The delimiting of ocean spaces—the traditional territorial sea and contiguous zone as well as the new extended economic and continental-shelf zones—was assigned to the powerful Second

ADOPTION OF CONVENTION on the Law of the Sea was photographed at United Nations Headquarters in New York on April 30 of last year. Above the delegates is a tally board that records the vote: 130 nations in favor, four against and 17 abstaining. At Montego Bay in Jamaica on December 6, 119 of the nations signed the Convention and 143 signed the Final Act of the Conference on the Law of the Sea. Signing the Final Act but not the Convention entitles a nation to attend future meetings of the Preparatory Commission as a nonvoting observer.

Committee. Here the conflicting interests of three groups of nations had to be reconciled. The coastal states, of course, sought to extend the limits of their jurisdiction as far as possible and to secure maximum attainable rights in the zone under their jurisdiction. At the other end of the spectrum were the landlocked states, including the numerous landlocked African nations, the poorest of the poor. They could claim no zones and stood to lose what rights they might have enjoyed, particularly with regard to fishing, in the seas about to be closed to them. Somewhere in the middle were the maritime states, sharing some of the interests of the coastal states but depending, on the other hand, on the freedom of the seas to deploy their distant-water fleets and so sharing interests with the landlocked states. Each of these groups was a complex aggregation of developed and developing countries with both free-market and centrally planned economies.

There can be no doubt that the largest gainers from the expansion of national jurisdiction were the industrial states, particularly the U.S., Canada, Australia and South Africa, and also a few big developing countries, India, Brazil and Mexico among them. The articles of the Convention issuing from the Second Committee affirm the 12-mile limit of the territorial sea as well as a 12-mile contiguous zone. Beyond these limits they accord to the coastal states sovereign rights over the use of natural resources in the newly sanctioned 200-mile economic zone and in addition a bundle of other rights with respect to such activities as scientific research. Where the continental shelf extends beyond 200 miles the economic-zone rights with respect to resources on the shelf are extended with it, but not beyond 350 miles.

The coastal state does not hold territorial sovereignty in its economic zone, however, as it does over its land and ad-



MARITIME CLAIMS of nations are compiled in this map based on one compiled by F. Webster McBryde of Transemantics, Inc. The areas in blue are fishing zones, where exclusive fishing rights are claimed. The areas in green are economic zones, where not only fishing rights are claimed but also sovereign rights over all resources and jurisdiction over research and marine pollution. The areas in orange are "territorial seas," where full sovereign rights are claimed subject to the right of "innocent passage" (transit without military, economic

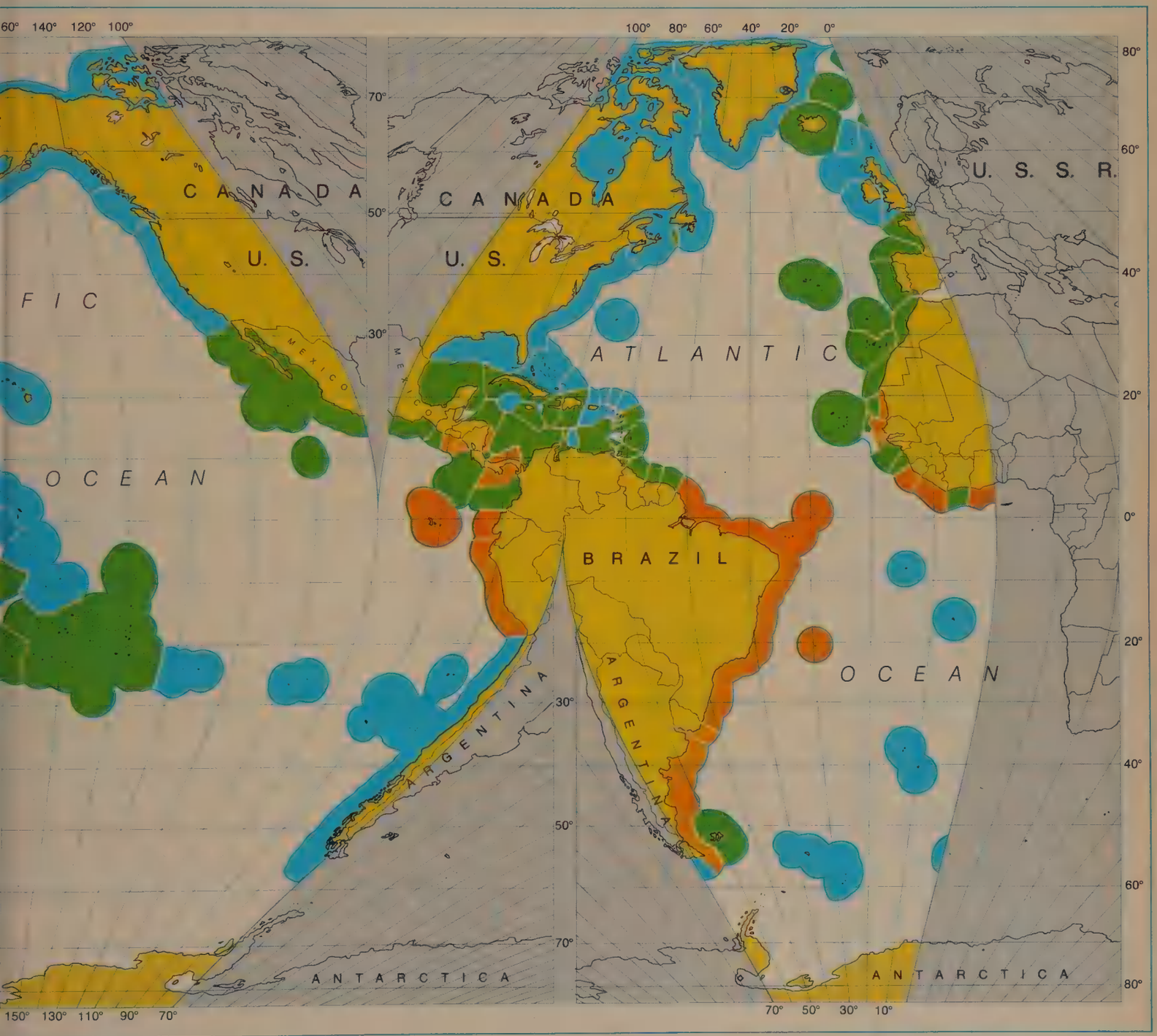
jacent territorial sea. This new kind of functional sovereignty may be regarded as the least malignant form of national aggrandizement. It is a new and dynamic concept in international law that undoubtedly will find other applications in the future.

The creation of the economic zone responds also to the need for management of the maritime environment. It provides an economic incentive for national action in this regard. It does not quite satisfy the need, however, since neither pollution nor fish respect national boundaries.

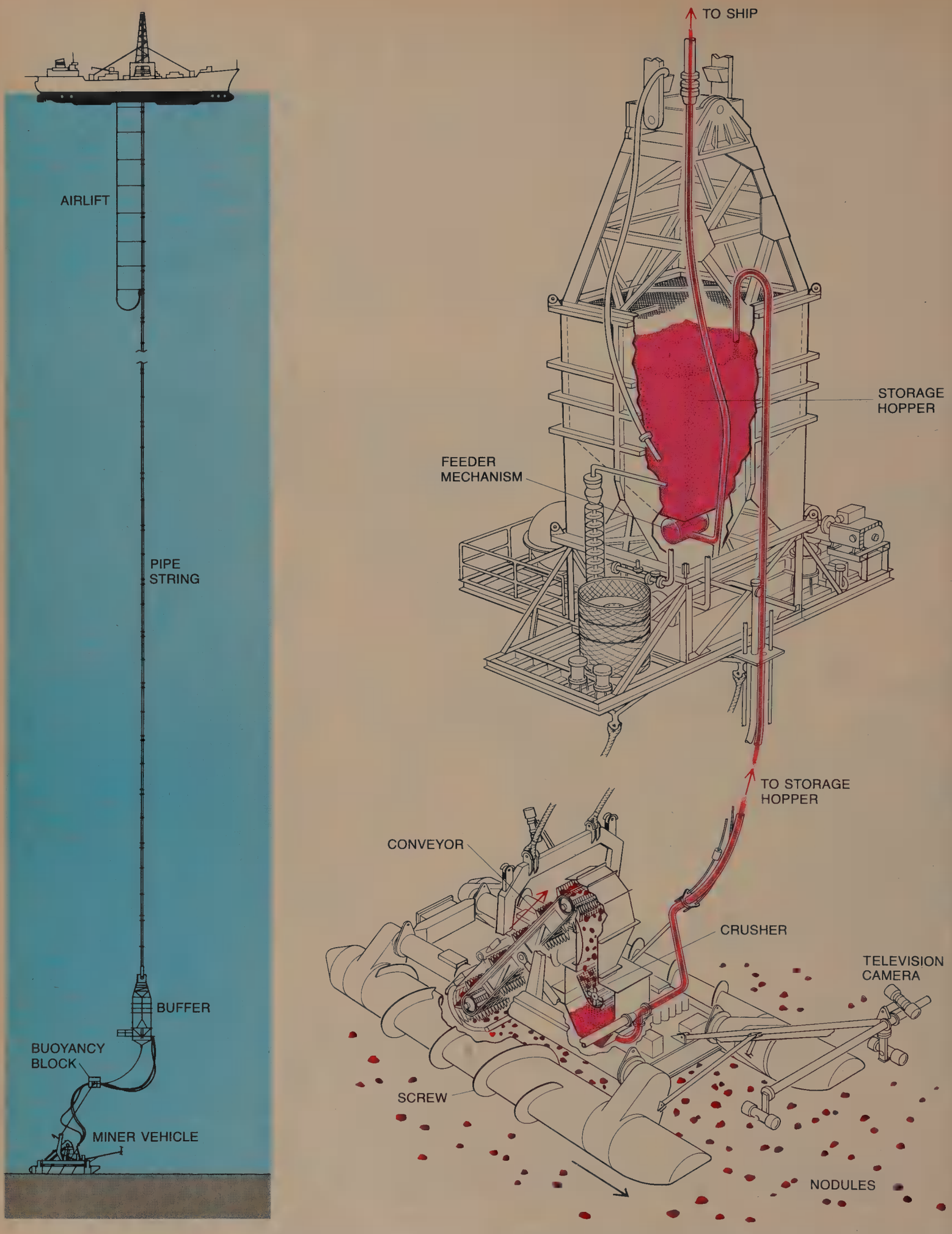
On the negative side the economic-zone concept has been called the greatest territorial grab in history. The expansion of national jurisdiction in the oceans surely increases inequality among nations. If it was hoped that the Conference would set clear and unambiguous boundaries to national jurisdiction, it must be said to have failed in this respect as well. The limits as defined are elastic, with loopholes for further claims at any time. It is an easy prediction that the discovery of any substantial resource anywhere in the world ocean will invite a claim from the nearest coastal or island state, in spite of the solemn declaration of the nonappropriability by any state or person (a corporation, for example) of the Common Heritage of Mankind "beyond the limits of national jurisdiction."

It can nonetheless be hoped that developments in the future will overcome the weaknesses and mitigate the injustices inherent in this part of the Convention. Most likely that will happen through progress in regional cooperation and the strengthening of international organizations. As will be seen, there is reason to look forward to some form of revenue sharing from the exploitation of the economic zone.

Political alignments were somewhat different in the First Committee, charged with setting up the international regime for the management of the Com-



or scientific significance). These claims, adding up to some 40 percent of the oceans, are recognized by the Convention on the Law of the Sea. The remaining 60 percent of the area is reserved for the traditional freedom of the seas, and the resources of that area of the ocean floor are deeded to the Common Heritage of Mankind. The projection is one designed by McBryde to avoid the distortion of both land and sea areas. The map can be obtained for $5 ($6.50 postpaid) from Transemantics, Inc., 1828 L Street NW, Washington, D.C. 20036.



MANGANESE NODULES, which are strewn over large areas of the ocean floor, could be harvested in a number of ways. The system shown here has been designed by Conrad G. Welling and his colleagues at the Ocean Minerals Company. The nodules typically contain 25 percent manganese, 1.5 percent nickel, 1.2 percent copper and .2 percent cobalt. In this system a miner vehicle (*bottom right*) is propelled across the bottom by two large screws. The nodules are picked up from the bottom by an endless-belt rake, crushed and pumped as a slurry to a "buffer" (*top right*) suspended by a pipe string from the surface ship. The slurry is then pumped to the ship.

mon Heritage of Mankind. Here the divergence of interests between the industrial nations of the North and the developing nations of the South was clearly the polarizing agent. The South held that the adoption of the Declaration of Principles by the General Assembly of the United Nations in 1970 had given the Common Heritage of Mankind the sanction of international law. In its view the task of the First Committee was to create a strong Seabed Authority with comprehensive functions and powers. The Authority would have an operational arm, the Enterprise, which would itself explore and exploit the Common Heritage on behalf of all mankind and could compel mining companies to sell it the technology needed for the task. Reacting to these proposals, the North proposed an Authority empowered to issue licenses and collect a small tax or royalty on the exploitation of the Common Heritage, which for the rest could be utilized by the North as it pleased.

The compromise that resolved this conflict was long and hard to come by. The Enterprise system conceived by the South was held to be an impractical proposition by the North, secure in its possession of the necessary capital and technology. The licensing agency advanced by the North was unacceptable to the majority of the nations convened in the Conference, who viewed it as a violation of the principle of the Common Heritage of Mankind. The compromise proposed by Henry Kissinger combined the impractical Enterprise with the unacceptable licensing agency in a "parallel system," a combination which, by a kind of witches' arithmetic, was to result in a solution both practical and acceptable! The Conference never quite recovered from it.

In the course of the tedious negotiations to spell out this compromise the text defining the parallel system became ever more abstruse and remote from reality. Playing on the fear of nonratification by the U.S. Senate, the American negotiator Leigh Ratiner succeeded in getting every administrative and financial detail spelled out in advance and engraved in articles that could not be altered for 25 years—and this for an industry not yet in existence and whose future was admittedly quite uncertain. The South meanwhile had second thoughts about the desirability of exploiting the Common Heritage at all at this time. Declining commodity prices made the resource-exporting developing countries wary of a new source of competition for their native ores. Against the opposition of the North (with the exception of Canada) these countries became more interested in limiting production from the seabed than in managing it.

The value of the Common Heritage of Mankind, particularly of the manganese nodules, has been discounted by other developments. Mexico and Chile have found rich fields of nodules in their economic zones; France has nodules in Polynesian waters, and the U.S. can assert claims offshore from the Hawaiian and other Pacific islands to hundreds of thousands of square miles of ocean floor likely to have commercially exploitable nodules. The nodules themselves have lost glamor to the deposits of polymetallic sulfides containing copper and other metals in much higher concentrations, first encountered at volcanic vents in the Pacific rift near the Galápagos Islands and off the coasts of Oregon and Washington. Similar deposits are certain to be found elsewhere along the 40,000-mile, world-girdling submarine rifted ridge.

Yet the creation of the International Seabed Authority by the consensus of practically the entire world community (in which the consent of the U.S. must be counted, because it was given at the Conference and will, I am sure, be given again) must be reckoned as a breakthrough in international relations. Here is an international institution unprecedentedly empowered to regulate and act on the basis of the new principle of the Common Heritage of Mankind. Here is a first attempt at a global production policy with due regard to conservation of the environment. Here is an opening to industrial cooperation between the North and the South based not on aid but on sharing. The European Economic Community was a dream of a few "federalists" 50 years ago. The International Seabed Authority, a utopian dream of 20 years ago, is now a fact of international law. Something has been moving.

The definition of a new regime for marine scientific research, for the protection and conservation of the marine environment and for encouraging the transfer of technology was the task of the Third Committee. Political polarization was less marked in the deliberations of this committee and it completed its work earlier than the others.

The freedom to do scientific research, one of the implied freedoms of the sea, is plainly constrained by the expansion of national jurisdictions into the ocean. It should be noted, however, that under the "consent regime" provided in the Convention a coastal state is normally expected to give its consent to nationals of other countries who want to do research in its waters. (U.S. scientists who have been worried about bureaucratic delays in getting coastal-state consent should now be concerned by their country's rejection of the Convention, which may prevent their even applying for such consent, much less getting it.) The consent regime should have the effect of encouraging bilateral cooperation in research enterprises, adjusting them to the interests of the coastal state as well as those of the investigators. More than any legal instrument in the past the Convention encourages the internationalization of marine research, time and time again making cooperation mandatory and calling on the "competent international organizations." Projects undertaken under the sponsorship of such organizations are deemed to have the consent of the coastal state, provided only that the coastal state has participated in, and not opposed, the project within the organization.

The measures for the protection and conservation of the marine environment provided in the Convention cover seaborne pollution, pollution from seabed mining and oil production and pollution from land-based sources and the atmosphere. This comprehensive framework for an international code of law even foresees the possible alteration of the environment through the introduction by aquaculture of exotic species or through the application of new technologies. States are to be held liable for damage inflicted not only in waters under the jurisdiction of other states but also on the high seas. The appropriate organizations, particularly the Intergovernmental Maritime Organization and the United Nations Environment Programme, are enjoined to get on with the work in which they are already engaged and find their authority thereby reinforced.

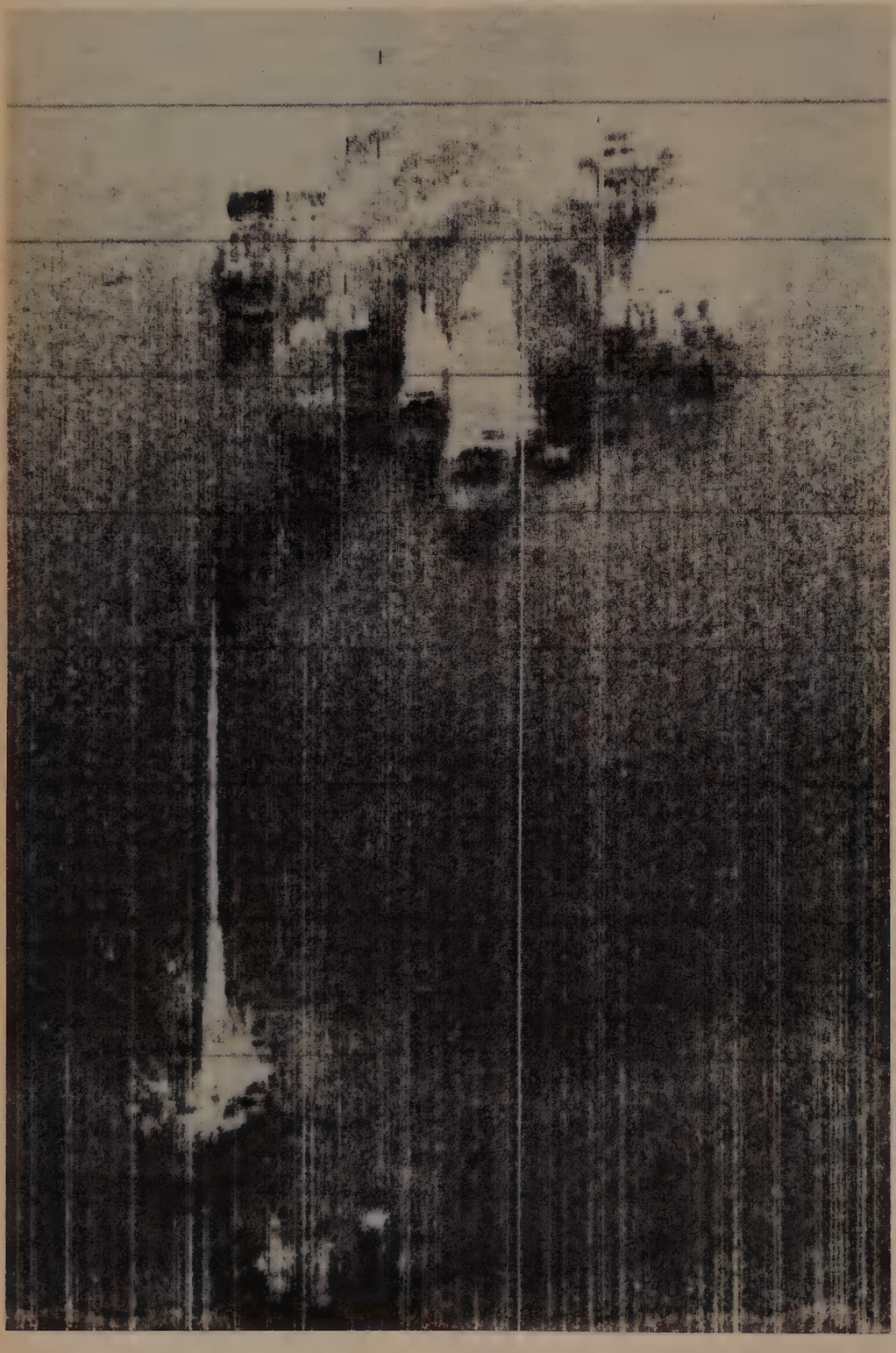
Indeed, through its Regional Seas Programme the United Nations Environment Programme has already begun the transformation of the "soft" law of the Convention into "hard" enforceable law. There are now 11 such programs engaging about 110 states and a great number of intergovernmental and nongovernmental organizations [*see illustration on page 49*]. Participating nations are taxing themselves for the costs of cleaning up the seas and preventing further pollution. While the Conference was still in session the United Nations Environment Programme was putting into force the universal standards and norms for pollution control set out in the draft Convention.

Encountering indifference rather than resistance from the North, the developing countries made the Convention articles on marine technology transfer their own. They went as far as they could, but the result is "soft" law indeed. The transfer of technology presents a complex technical, social, political and educational task. One may well ask whether it can be legislated. The law does what it can to set goals and specify modes of cooperation; the rest is up to history. That history is already delineating itself. In the industrial countries the superannuated patent system has small correspondence to the realities of technolog-

ical competition. Intellectual property is bound to find its natural place in the Common Heritage of Mankind.

The Convention makes an important contribution to the development of a binding dispute-settlement system and thus to the maintenance of peace. It establishes, in fact, the most comprehensive and binding global system ever devised. A new supreme court is established in the International Tribunal of the Law of the Sea that is to have its seat, appropriately, in the Hanseatic city of Hamburg. In one of its special chambers, the Seabed Chamber, corporations and individuals will have standing. Regrettably, but inevitably, there are loopholes in the system. States can claim exemption from compulsory dispute-settlement procedures, particularly with regard to the most sensitive issues, for example the delimitation of sea boundaries or issues arising from the military uses of the sea.

In this, as in so many other respects noted here, the Convention is an ambiguous document. The Law of the Sea does not, however, lead an autonomous existence. It must be seen as the expression of the order and disorder, intranational as well as international, of our time. If present trends continue in other sectors of international relations, states will exploit the loopholes, defects and contradictions of the Convention to plunder and pollute the ocean, exhaust its living resources and deploy their weapons in its depths. If states, on the other hand, turn to serious negotiation of balanced global development and peace, they will find that the Convention provides a foundation for building a better world.



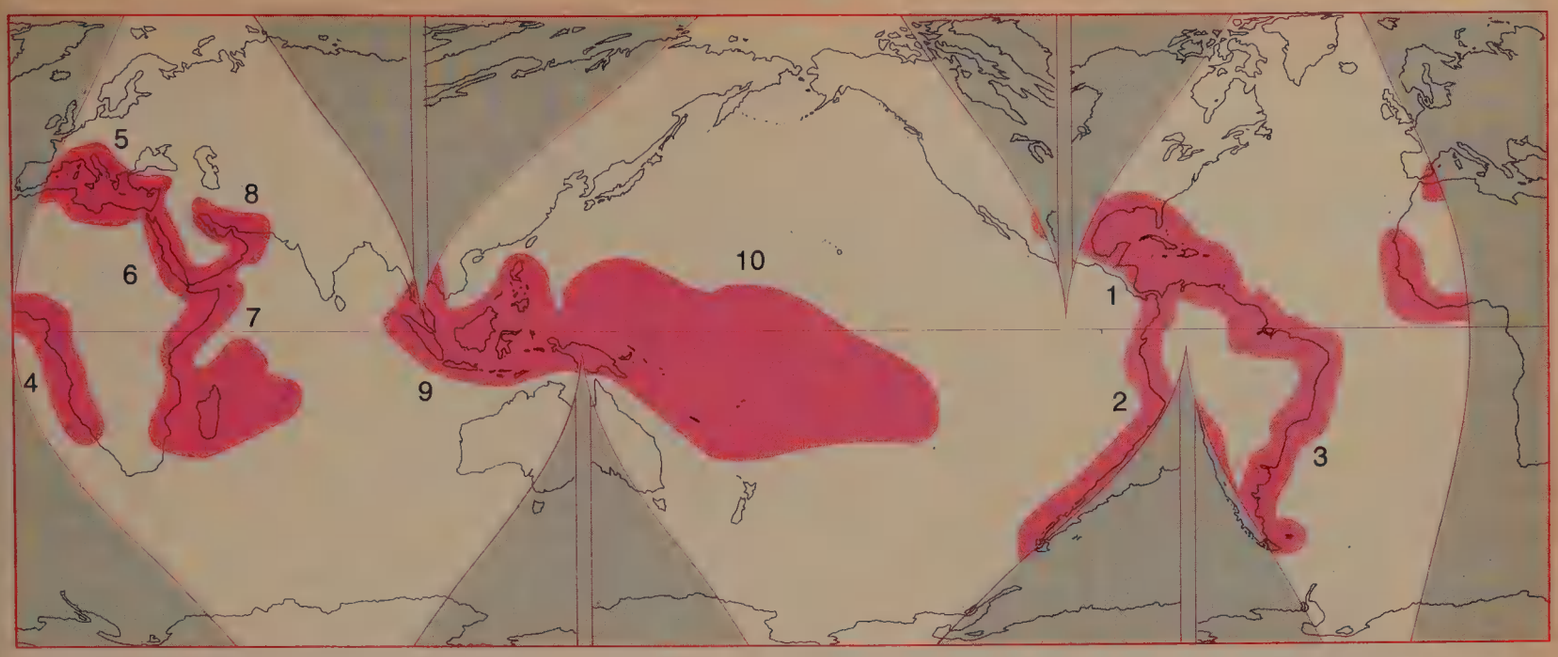
DEEP-SEA MINERAL DEPOSITS in the Gulf of California are revealed by this side-looking sonar image. The features projecting up from the bottom are mounds and "spires" characteristic of hot-spring areas adjacent to sea-floor-spreading centers. In such areas metal-rich sulfides precipitate out of hot water flowing up through the bottom. The features extend 30 meters above the bottom. The distance between four horizontal lines is 450 meters. The image was supplied by Fred N. Spiess and Peter F. Lonsdale of the Scripps Institution of Oceanography.

The ocean is bound to play an ever greater role in the world economy. From small present beginnings in aquaculture the traditional hunting-and-gathering technology will give way to the cultivation of aquatic plants and the husbandry of aquatic animals. The seabed will meet the demand for metals long after the thinning of continental ores has exhausted the ingenuity of technology. By the end of the century the seabed will supply more than half of the world's oil and the ocean will hold energy in reserve in its tides, waves and currents, in its thermal and salinity gradients, in its biomass and prospectively not least in the rare but abundant deuterium atoms of its water molecules.

The ascending economic importance of the ocean will bring it more directly into the political life of nations and of the international community. Nations are creating departments of ocean development, ministries of ocean affairs and the like to assume new functions of ocean management. They are updating their maritime law and extending the rule of law into areas and activities hitherto not covered by law. Harmonizing their practices and institutions with the Convention on the Law of the Sea, they are internalizing and implementing the provisions of the Convention.

The competent international organizations so often invoked in the Convention—the Intergovernmental Maritime Organization, the Food and Agriculture Organization of the United Nations, the Intergovernmental Oceanographic Commission of UNESCO and the United Nations Environment Programme—are taking up their tasks. Particularly for coastal developing countries these agencies are charged with helping to frame national legislation, to improve the management of living resources in their extended economic zones, to chart shipping lanes and schemes for traffic control, to set up regional centers for the advancement of ocean science and technology and to arbitrate conflicts.

Although it seems paradoxical, the expansion of national jurisdiction into the ocean does not diminish the need



"ACTION PLANS" are in effect or in various stages of preparation under the Regional Seas Programme of the United Nations Environment Programme. Plans in effect are those for (*5*) the Mediterranean, (*6*) the Red Sea and the Gulf of Aden and (*8*) the Kuwait Action Plan Region. Plans in preparation are those for (*1*) the Caribbean, (*2*) the Southeast Pacific, (*3*) the Southwest Atlantic, (*4*) West Africa, (*7*) East Africa, (*9*) the East Asian Seas and (*10*) the Southwest Pacific. The projection is the same as that of the map on pages 44 and 45.

for international cooperation; on the contrary, it increases the need greatly. There can be no international cooperation without well-organized nations. In the technological and economic reality of today nations cannot maintain their economic health without strong international cooperation.

The worldwide networks of the Regional Seas Programme are breaking paths to new modes of international cooperation. So important to each nation are its interests in the sea that Turks and Greeks, Israelis and Arabs, Iraqis and Iranians have continued cooperation in their Regional Seas programs even in times of war. Success of a Regional Seas Programme in the Indian Ocean could give impetus to the Sea of Peace concept for that region, which has been adopted and reiterated by the UN but not respected by the superpowers. In the Mediterranean the development of machinery for the funding of the Regional Seas Programme might revive an idea first put forward by Malta and Mexico in the deliberations of the Seabed Committee in 1972: that nations be taxed on their use of the ocean, on revenues from offshore oil, on commercial fisheries and on seaborne trade.

The signing of the Convention at Montego Bay brings into action as of March a new organization, the Preparatory Commission, charged with preparing the rules and regulations for the International Seabed Authority and a provisional agenda for the first session of its governing bodies. The convening of the first session, however, must await the ratification of the Convention. In the interim, the duration of which cannot be predicted, the Preparatory Commission is endowed by the Conference with some of the powers of the Authority itself. Among its considerable powers, functions and responsibilities, the Commission is to examine applications from "pioneer investors" for exploration sites; verify prospecting data and choose the sites to be reserved for the future Authority and its Enterprise, and negotiate arrangements for technology transfer to the Authority and for the training of personnel from developing countries. In sum, the Preparatory Commission has considerable operational capacity. In a way it provides an interim regime for exploration, research and development in ocean mining technology, and it is likely that activities of the Seabed Authority itself will be limited to exploration, research and development until the end of this century.

Once it has acquired the requisite technology—possibly through a joint venture or joint ventures with mining companies—the Commission could apply it not only to explore for nodules in the international area but also to help developing countries that invite such assistance to prospect and develop the resources of their national economic zones. There is nothing in the Convention that would prevent the Authority, or the Commission before it, from performing such needed and useful services, provided they are sought by a coastal state.

Another immediately useful service the Commission could render would be to create marine parks, regions "disapproved for exploitation," and reserve them for scientific purposes. The first such park could well cordon off the Galápagos vents and the unearthly colonies of giant worms and crustaceans sustained there by bacterial chemosynthesis that evolves sulfur rather than oxygen, as in the more familiar and ubiquitous life-sustaining photosynthesis.

A third immediately useful service would relate to arms control and disarmament. The treaty of 1972 banning nuclear weapons from the sea floor made no provision for inspection to verify the observance of its terms. Inspection was left by default to the superpowers. The majority of the countries wanted to internationalize the monitoring and surveillance but were frustrated by the lack of a competent international organization. The Preparatory Commission, and the Authority, could fill the need.

By 1987, when the seabed disarmament treaty next comes up for review, the Commission may well have acquired just the know-how needed; the same acoustic, seismic, magnetic and electronic technologies employed in resource exploration can be turned to the inspection task. An amendment to the treaty in 1987, which would be favored by a large majority of the signatories, could entrust this task to the Preparatory Commission or to the Authority itself if the Convention has been ratified by then.

The signing of the Convention at Montego Bay by the overwhelming majority of the international community thus opens a new phase of the building of an international order. It offers new instruments for enhancing economic development, advancing disarmament, ensuring protection of the environment and creating new forms of scientific and industrial cooperation between North and South. These possibilities would not be there without the new Law of the Sea.

Microprogramming

In most modern computers the routing of information is controlled at the lowest level by a microprogram: a set of stored instructions that functions in place of a completely "hard-wired" control system

by David A. Patterson

A fundamental issue in the design of any computer is how to control, or steer, the electrical signals that represent information. In the arithmetic and logic unit, where the actual processing of information is done, signals must be routed between various counters, adders and other components. The control system must also mediate the transfer of information between the central processor, the main memory units and the various input and output devices. In one approach the control system is completely "hard-wired," that is, it is laid down permanently in the processor's electrical circuitry. A second approach is more flexible and in many cases less expensive. The essential idea is to reduce the complexity of the control system by recording the detailed instructions for controlling the computer in a coded form. In other words, the sequence of paths a signal is to follow is embodied in a program, which is stored in a separate memory unit incorporated into the processor.

In the hierarchy of programs that operate a computer the instructions executed by the control system occupy the lowest and most elementary level; each instruction specifies a single functional state of the machine. Because the control instructions are responsible for such fine details, the task of defining and encoding them is termed microprogramming, thereby distinguishing it from the writing of the higher-level programs known generically as software. A set of control instructions—a microprogram—is said to be written in microcode. The idea of microprogramming was conceived more than 30 years ago, soon after the advent of the first computers. At the time the hardware needed to implement the idea did not yet exist. The method has been adopted, however, in most computers being built today.

In recent years a good deal of confusion has arisen about the meaning of the term microprogramming, owing largely to the advent of the microprocessor, the "computer on a chip" that is at the heart of the latest products of the progressive miniaturization of silicon-based semiconductor technology. It must be emphasized that microprogramming a computer is not the same as programming a microcomputer; in principle any computer, from the largest "mainframe" system to the smallest personal computer, can be designed with a microprogrammed control system. To avoid such confusion microprograms are sometimes classified as firmware, thereby signifying their intermediate status between hardware and software. For clarity I shall adopt the following convention: the term macroinstruction will refer to a software instruction and the term microinstruction will refer to a firmware instruction.

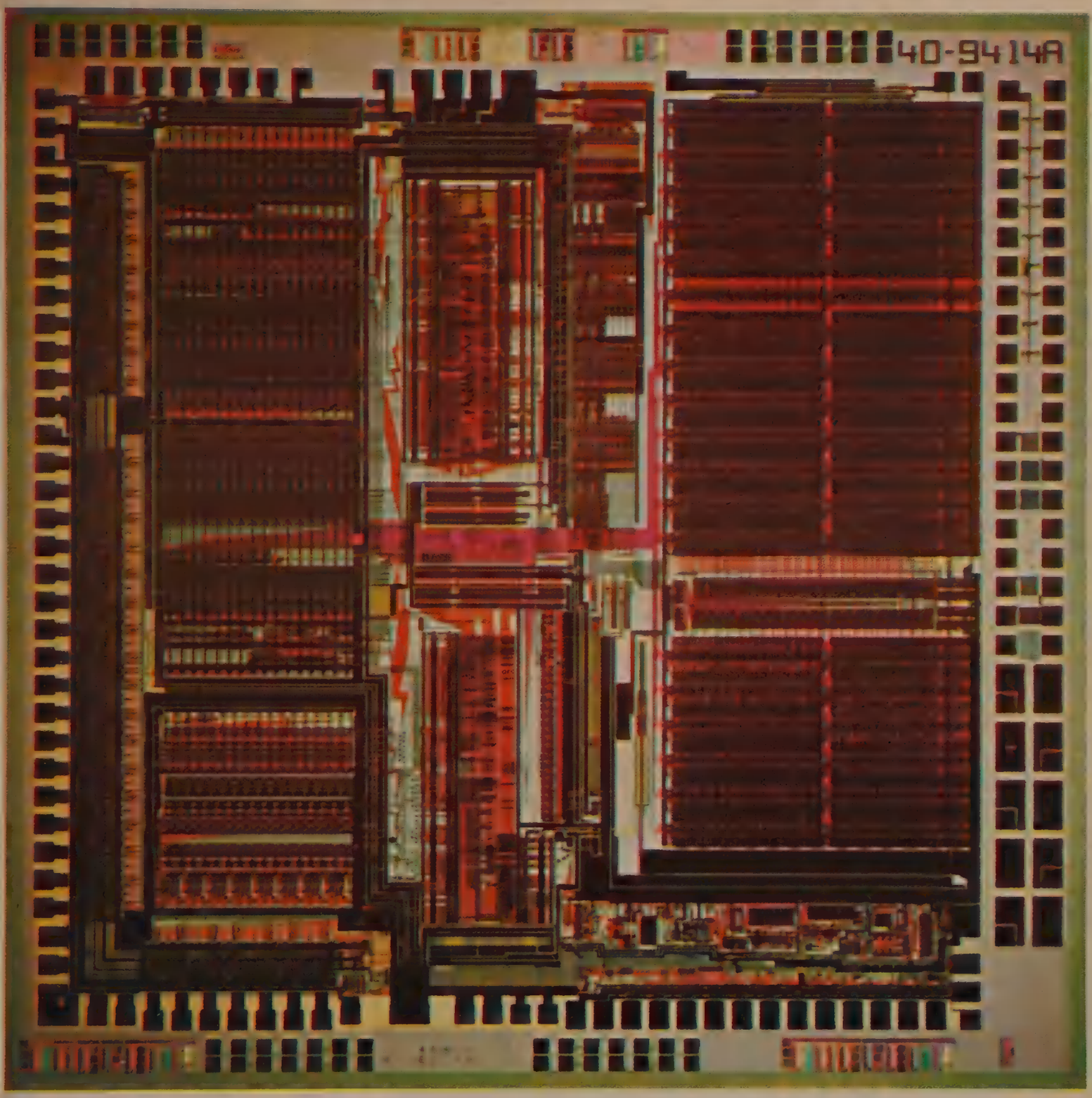
Before microprogramming can be explained it is necessary to review some of the basic principles of computer hardware and computer programming. In many ways a computer is like a simple calculator. Numbers are introduced into both devices by means of a keyboard, and they are read out on a display unit. Some calculators have a memory key to record a number for later retrieval; saving and restoring numbers is also the purpose of the memory unit in a computer. A computer memory, however, is capable of handling millions of numbers. To keep track of so many numbers each one is assigned a specific address.

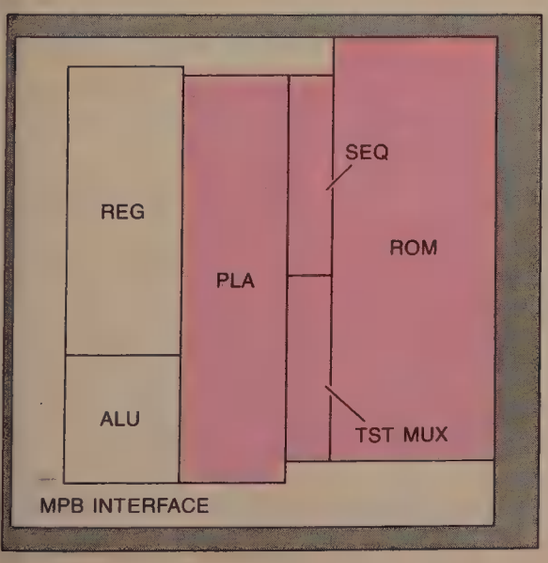
Both a calculator and a computer must be told what to do. Giving a computer a command is analogous to pressing an operator key on a calculator. For example, almost every computer has the macroinstructions ADD, SUB, MULT and DIV, which correspond exactly to the keys marked $+$, $-$, $\times$ and $\div$ on a calculator. Similarly, the macroinstruction LOAD performs the same function as the memory key on a calculator.

Unlike a simple calculator, however, a computer is preeminently a decision-making machine. To understand the distinction consider the sequence of operations represented by the last few sections of Form 1040, the U.S. Individual Income Tax Return [*see top illustration on page 53*]. A computer programmer faced with the chore of completing the form might begin by making a flow chart, which serves as a kind of blueprint for constructing a program [*see illustration at bottom left on page 53*]. A taxpayer working with a simple calculator could follow the flow chart, carrying out the instructions in sequence until he reached the diamond-shaped "decision" step. He would then compare the numbers in the preceding two boxes and, depending on which number was greater, choose the appropriate path to continue his calculations.

For a taxpayer working with a computer rather than a calculator the programmer could go a step further: he could get the computer to make the decision by itself. Following the flow chart, he would first instruct the computer to add successively the numbers entered on lines 50 through 58 of the tax form, storing the sum of each addition at a specific address in the computer's memory; in this case the addresses of the numbers are designated $L50$, $L51$, $L52$ and so on, and they are added successively to an "accumulator" address designated A [*see illustration at bottom right on page 53*]. The total for this sequence of steps would then be stored at address $L59$. Next the programmer would do the same for lines 60 through 66, storing the total for the second sequence at address $L67$. At this point (step 19) he would give the computer the following macroinstruction: "JUMPIF $A < L59$, 23." This macroinstruction, called a conditional jump, tells the computer to begin executing a new sequence of macroinstructions at step 23 if the number at the accumulator address is less than the number at address $L59$. If it is not, the program would continue with step 20.

It is the conditional-jump macroinstruction, corresponding to the diamond-shaped box in the flow chart, that endows the computer with the ability to





MICROPROGRAMMED MICROPROCESSOR is at the heart of the Hewlett-Packard Company's HP 9000 family of computers. The microelectronic circuit is the densest of its kind ever made: it has the equivalent of some 450,000 transistors on a square chip of semiconducting silicon 6.3 millimeters on a side. The standard spacing between the signal-carrying lines is one micrometer. The map at the left shows the seven major subsections of the chip. The colored area identifies those subsections that constitute the processor's control system and the white area corresponds to the data path. The control system consists of four parts: a read-only memory (*ROM*), where the microinstructions for controlling the flow of information throughout the computer are stored; a programmed-logic array (*PLA*), where most of the microinstructions are decoded before being transmitted to their various destinations; a sequencing device (*SEQ*), where the order of the microinstructions is arranged, and a test multiplexer (*TST MUX*), where certain conditional microinstructions are decoded. The three other subsections are an array of memory registers (*REG*), where the data being handled by the computer are stored in the form of "words," each of which is 32 bits long; an arithmetic and logic unit (*ALU*), where the actual processing of the 32-bit words is done, and a memory-processor-bus (*MPB*) interface, which oversees communications between the processor chip and the rest of the computer. The rectangular shapes at the top, bottom and right edges of the chip include test devices, special resolution patterns and alignment marks that are unrelated to the operation of the microprocessor.

make decisions, provided of course that the programmer can represent the decisions as simple arithmetic tests. The programmer builds a program by constructing a maze of conditional jumps that leads each input number to its correct destination. On the average one out of 10 macroinstructions in a computer program is a conditional-jump macroinstruction; many computers available today can execute a macroinstruction in less than a millionth of a second, so that a computer typically makes more than 100,000 decisions of this kind per second. The macroinstructions for making the decisions are themselves encoded as numbers and stored along with the other numbers handled by the computer in the main memory. The principle of the stored program, the invention of which was a milestone in the development of the modern digital computer, makes it possible to change the function of a computer by changing the contents of its memory unit instead of by changing its hardware.

As computers get faster they can execute more macroinstructions and make more decisions in a given time span; hence it becomes possible to run increasingly complex programs. Under the circumstances mistakes are difficult to avoid. Programmers prefer a term whose connotations are not quite the same as those of "mistake," however: they say that a program has a "bug." To appreciate the special sense of this usage, consider again the illustration of the income-tax form on the opposite page. Line 68 tells the taxpayer what to do if the amount on line 67 is larger than the amount on line 59, and line 71 explains what to do if line 59 is larger than line 67, but nowhere in the form or in the accompanying instructions is there any indication of what to do if the amounts on lines 59 and 67 are equal! Evidently the Internal Revenue Service has overlooked a bug in its program. It is precisely this kind of detail that a programmer must keep in mind when he is writing and "debugging" a computer program. The difficulty of the task is compounded when it comes to the fine details of writing and debugging a microprogram.

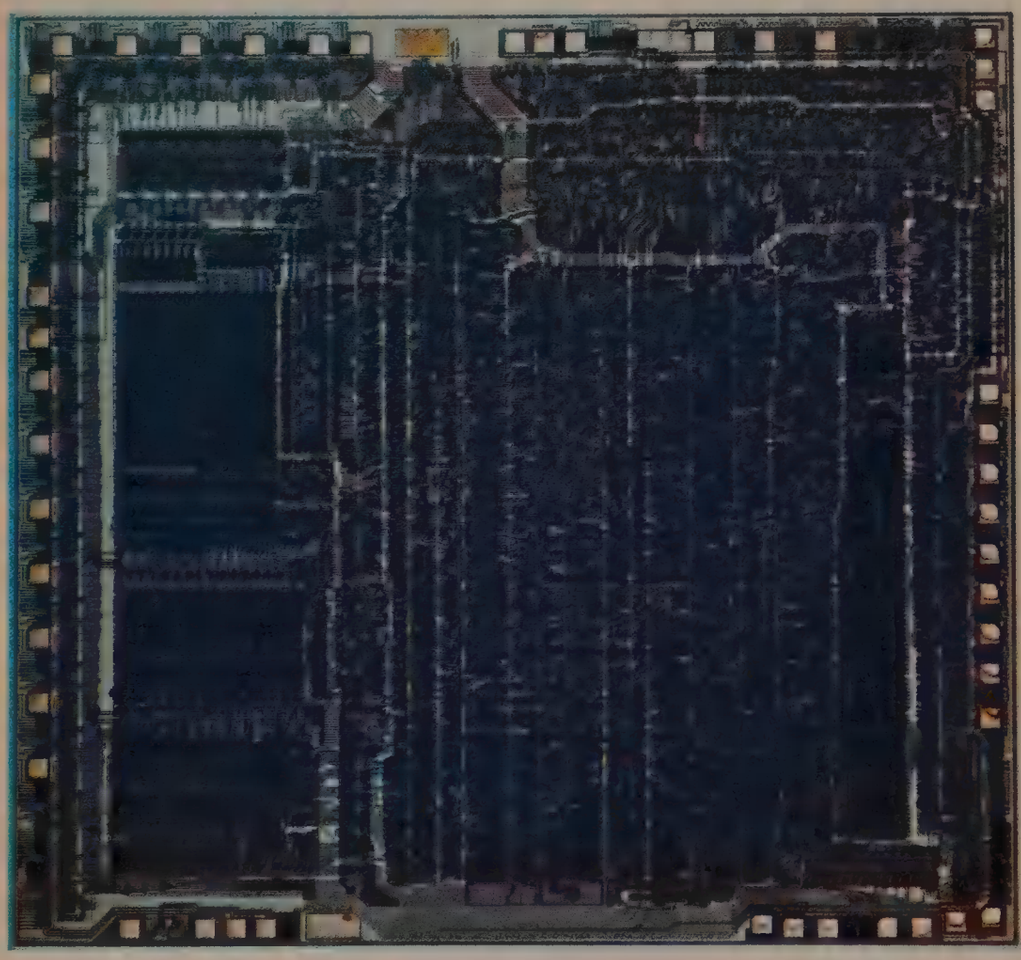
To see how decisions are made and macroinstructions are executed by a computer one must look deeper into the operation of the central processing unit. The processor consists of two major parts: the control system, which makes the decisions and issues the resulting commands to the rest of the hardware, and the arithmetic and logic unit, or data path, where calculations are actually carried out. At the heart of the arithmetic and logic unit is an adder, and in principle all the arithmetic functions can be reduced to repetitive addition steps. Each part of the processor has its own array of short-term memory devices, called registers.

The organization of the computer at this level resembles plumbing. The wires connecting the hardware devices can be thought of as pipes channeling the flow of information. The control system directs the flow by sending an "on" or an "off" signal to open or close each of a large number of valvelike electronic devices called gates [*see top illustration on page 54*]. Because the gates must be opened and closed at the right moment to get the information to the right place, the control signals must be precisely timed. The rhythm of the entire processor is therefore synchronized to an internal clock. Each macroinstruction corresponds to a sequence of clock cycles, with each clock cycle marking a single transfer of information along the data path. The function of the control system is to supply the control signals during the right clock cycles.

In a hard-wired control system a network of electronic logic must be devised to recognize each macroinstruction in the computer's repertory. The macroinstruction itself is a pattern of binary digits, embodied in a pattern of signals sent to the processor. The control system must somehow transform this pattern into the quite different set of signals needed to open and close various gates in the data path. Furthermore, the signals issued by the control system must be timed to coincide with the appropriate clock cycles. Each macroinstruction must give rise to a different sequence of control signals, and so the system becomes more complex as the number of macroinstructions increases. Changes can be made and bugs can be eliminated only by rewiring.

In the early days of the computer the design of the control hardware was a major problem. In comparison the task of designing the other components of the hardware was fairly easy. A teletypewriter could serve as an input-and-output device, and the repetitiveness of the memory structures and of the circuit elements in the data path simplified their design and construction. The hardware needed to count clock cycles and to issue control signals at the appropriate times made the control system quite irregular, and that in turn made the job of designing this part of the central processor much harder.

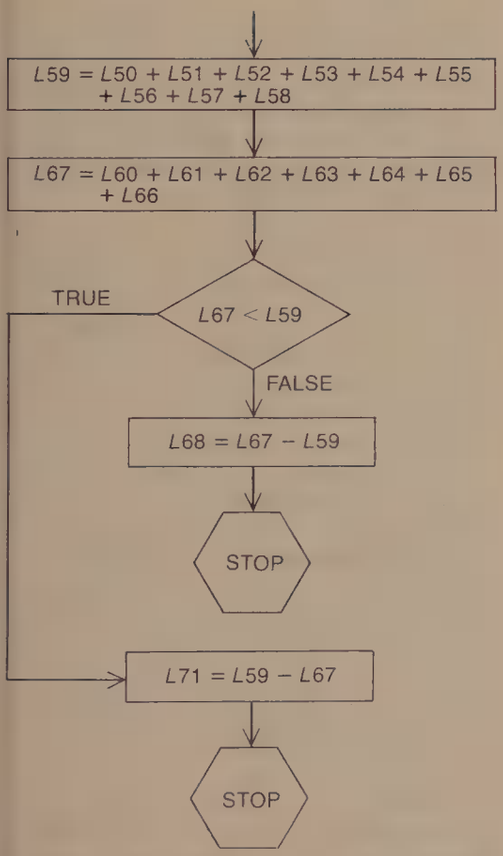
In 1949 Maurice V. Wilkes of the University of Cambridge set out to find a better way to organize the control function. After two years of study he concluded that the best approach is to



NONMICROPROGRAMMED MICROPROCESSOR is distinguished by its highly irregular, hard-wired control section, which occupies the large area at the right in this photograph. The two smaller and more regular areas at the left correspond to the *REG* and *ALU* subsections in the illustration on the preceding page. The irregularity of such a hard-wired control system makes it difficult to design and modify. At the time this chip was made (about five years ago) the designers thought that microprogramming would have required a much larger chip.

	50	**Balance.** Subtract line 49 from line 40 and enter difference (but not less than zero) ▶				50	
Other Taxes (Including Advance EIC Payments)	51	Self-employment tax (*attach Schedule SE*)				51	
	52	Minimum tax (*attach Form 4625*)				52	
	53	Alternative minimum tax (*attach Form 6251*)				53	
	54	Tax from recapture of investment credit (*attach Form 4255*)				54	
	55	Social security (FICA) tax on tip income not reported to employer (*attach Form 4137*)				55	
	56	Uncollected employee FICA and RRTA tax on tips (*from Form W–2*)				56	
	57	Tax on an IRA (*attach Form 5329*)				57	
	58	Advance earned income credit (EIC) payments received (*from Form W–2*)				58	
05	59	**Total tax.** Add lines 50 through 58 ■				59	
Payments Attach Forms W–2, W–2G, and W–2P to front.	60	Total Federal income tax withheld	60				
	61	1982 estimated tax payments and amount applied from 1981 return	61				
	62	Earned income credit. If line 33 is under $10,000, see page 15 of Instructions	62				
	63	Amount paid with Form 4868	63				
	64	Excess FICA and RRTA tax withheld (two or more employers)	64				
	65	Credit for Federal tax on special fuels and oils (*attach Form 4136*)	65				
	66	Regulated Investment Company credit (*attach Form 2439*)	66				
	67	**Total.** Add lines 60 through 66 ▶				67	
Refund or Amount You Owe	68	If line 67 is larger than line 59, enter amount **OVERPAID** ▶				68	
	69	Amount of line 68 to be **REFUNDED TO YOU** ▶				69	
	70	Amount of line 68 to be applied to your 1983 estimated tax ▶	70				
	71	If line 59 is larger than line 67, enter **AMOUNT YOU OWE.** Attach check or money order for full amount payable to Internal Revenue Service. Write your social security number and "1982 Form 1040" on it. ▶ (Check ▶ ☐ if Form 2210 (2210F) is attached. See page 16 of Instructions.) ▶ $				71	

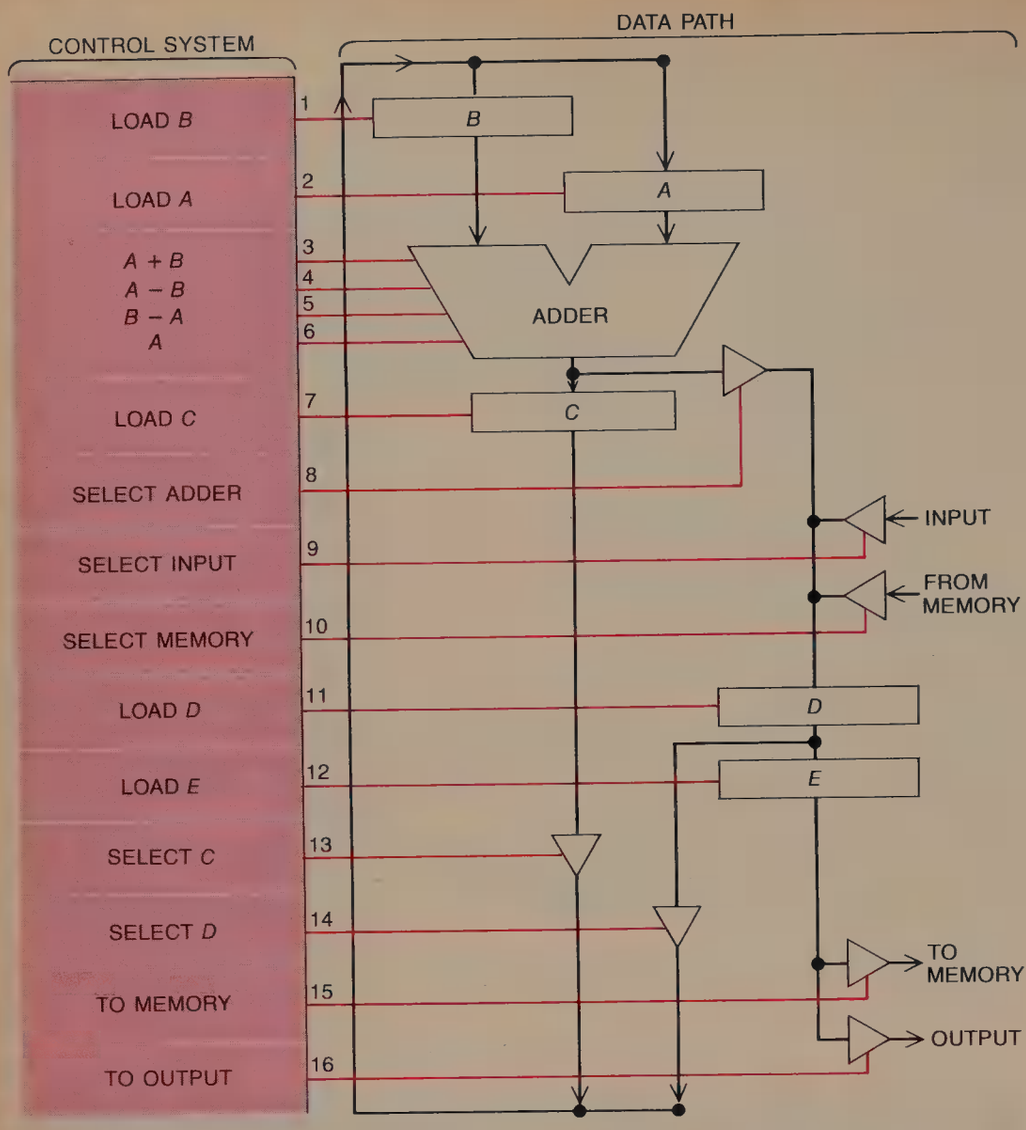
FAMILIAR FORM serves as ■ example of ■ sequence of operations that ■ be done by a computer program. The illustration reproduces the last 22 lines of Form 1040, the U.S. Individual Income Tax Return for 1982, prepared by the Internal Revenue Service.



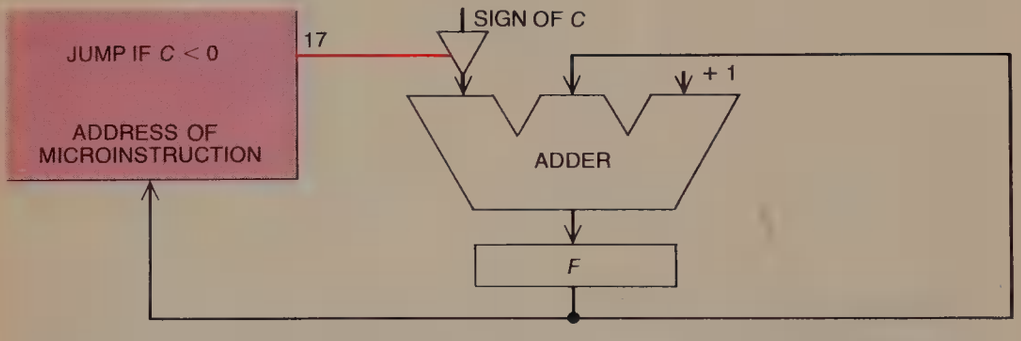
ADDRESSES	INSTRUCTIONS	COMMENTS
1	LOAD *L*50, *A*	*A* = *L*50
2	ADD *L*51, *A*	*A* = *L*50 + *L*51
3	ADD *L*52, *A*	*A* = *L*50 + *L*51 + *L*52
4	ADD *L*53, *A*	*A* = *L*50 + *L*51 + *L*52 + *L*53
5	ADD *L*54, *A*	*A* = *L*50 + *L*51 + *L*52 + *L*53 + *L*54
6	ADD *L*55, *A*	*A* = *L*50 + *L*51 + *L*52 + *L*53 + *L*54 + *L*55
7	ADD *L*56, *A*	*A* = *L*50 + *L*51 + *L*52 + *L*53 + *L*54 + *L*55 + *L*56
8	ADD *L*57, *A*	*A* = *L*50 + *L*51 + *L*52 + *L*53 + *L*54 + *L*55 + *L*56 + *L*57
9	ADD *L*58, *A*	*A* = *L*50 + *L*51 + *L*52 + *L*53 + *L*54 + *L*55 + *L*56 + *L*57 + *L*58
10	STORE *A*, *L*59	*L*59 = *L*50 + *L*51 + *L*52 + *L*53 + *L*54 + *L*55 + *L*56 + *L*57 + *L*58
11	LOAD *L*60, *A*	*A* = *L*60
12	ADD *L*61, *A*	*A* = *L*60 + *L*61
13	ADD *L*62, *A*	*A* = *L*60 + *L*61 + *L*62
14	ADD *L*63, *A*	*A* = *L*60 + *L*61 + *L*62 + *L*63
15	ADD *L*64, *A*	*A* = *L*60 + *L*61 + *L*62 + *L*63 + *L*64
16	ADD *L*65, *A*	*A* = *L*60 + *L*61 + *L*62 + *L*63 + *L*64 + *L*65
17	ADD *L*66, *A*	*A* = *L*60 + *L*61 + *L*62 + *L*63 + *L*64 + *L*65 + *L*66
18	STORE *A*, *L*67	*L*67 = *L*60 + *L*61 + *L*62 + *L*63 + *L*64 + *L*65 + *L*66
19	JUMPIF *A*< *L*59, 23	GO TO ADDRESS 23 IF *A* < *L*59 (SAME AS *L*67 < *L*59)
20	SUB *A*, *L*59	(*L*67 ≥ *L*59) THEREFORE *L*67 − *L*59
21	STORE *A*, *L*68	*L*68 = *L*67 − *L*59
22	STOP	FINISHED
23	LOAD *L*59. *A*	(*L*67 < *L*59) THEREFORE *L*59 − *L*67
24	SUB *L*67, *A*	*A* = *L*59 − *L*67
25	STORE *A*, *L*71	*L*71 = *A* (= *L*59 − *L*67)
26	STOP	FINISHED

FLOW CHART serves ■ ■ kind of blueprint of the sequence of operations needed to complete the last 22 lines of the income-tax form. The amount entered ■ each line of the form is stored at ■ specific address in the computer's memory; in this case the addresses of the numbers are designated *L*50, *L*51, *L*52 and so on. The diamond-shaped box is computer programmer's symbol for ■ "decision" step.

PROGRAM OF MACROINSTRUCTIONS for completing the income-tax form follows the flow chart. Each macroinstruction is itself stored at ■ memory address (*first column at left*). The numbers at addresses *L*50, *L*51, *L*52 and so ■ are added successively to an "accumulator" address designated *A*. Steps 1 through 10 correspond to the topmost box in the flow chart and and steps 11 through 18 correspond to the second box from the top. The conditional-jump macroinstruction at step 19 ("JUMPIF *A* < *L*59, 23") is equivalent to the diamond-shaped decision box. Steps 20 and 21 correspond to the box on the "False" path of the flow chart and steps 23, 24 and 25 correspond to the box ■ the "True" path. The comments at right ■ notes to remind the programmer of what is going on; they do not affect the execution of the program.



IDEALIZED CENTRAL PROCESSOR relies on 16 control wires (*colored lines*) to guide the electrical signals representing information through a simplified data path (*black lines*). In this case the arithmetic and logic unit has only one special-purpose device: an adder, which receives numbers from two input wires, adds or subtracts them in accordance with commands received over control wires *3* through *6* and sends the result through an output wire. The rectangular shapes are registers, short-term memory components that can hold one word of information at a time. The triangular shapes are gates, devices that pass the information along when the associated control wire is on and block the information when the control wire is off. In a nonmicroprogrammed processor each of the commands listed in the control section is defined by specific control hardware. In the microprogrammed version of the same processor each command is stored as a separate microinstruction in the microprogram memory. A sample microprogram for this simplified processor is given in the top illustration on the opposite page.



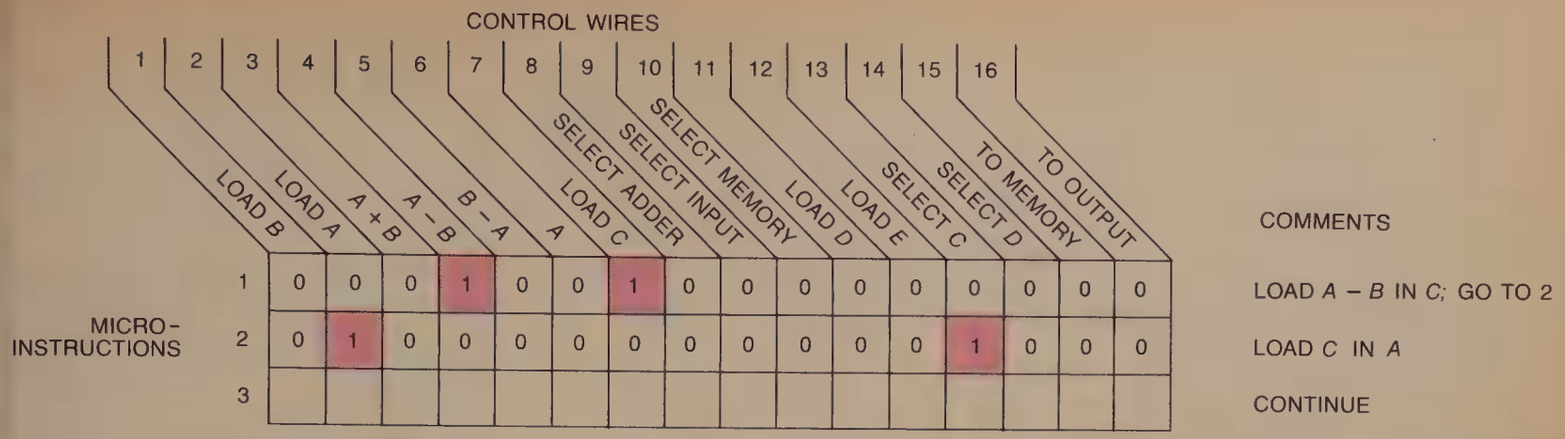
ADDITIONAL HARDWARE IS NEEDED to enable the microprogrammed version of the simplified processor depicted at the top of the page to make decisions of the conditional-jump type. Register *F* holds the address in microprogram memory of the microinstruction currently in control of the data path. The function of the extra adder is to add one ("+1") to the address of the current microinstruction, thereby transmitting the address of the next sequential microinstruction in the microprogram. The adder also feeds information from the data path back into the control system. The sign of the number in register *C* goes through a gate controlled by wire *17* into the third input of the adder. When this wire is off, the microinstruction at $F+1$, the next sequential address, is executed (since $0+F+1$ is equivalent to $F+1$). The adder adds either $0+F+1$ or $1+F+1$ depending on whether number in register *C* is respectively positive or negative. Thus with the aid of control wire *17* the microprogrammer can conditionally jump over a microinstruction, depending on the sign of the number in *C*. The microprogram for the conditional-jump operation is given in the bottom illustration on the opposite page.

think of the control system as a matrix, or rectangular grid, with each row of squares corresponding to a clock cycle and each column associated with a control wire. For example, the simplified processor shown in the top illustration at the left has 16 control wires, and so the control matrix corresponding to it is 16 squares wide. Choosing a sequence of operations then becomes a simple matter of putting a suitable combination of binary symbols in the squares: for each control wire that must be on during a particular clock cycle write a 1 in the corresponding square, and for each control wire that must be off write a 0.

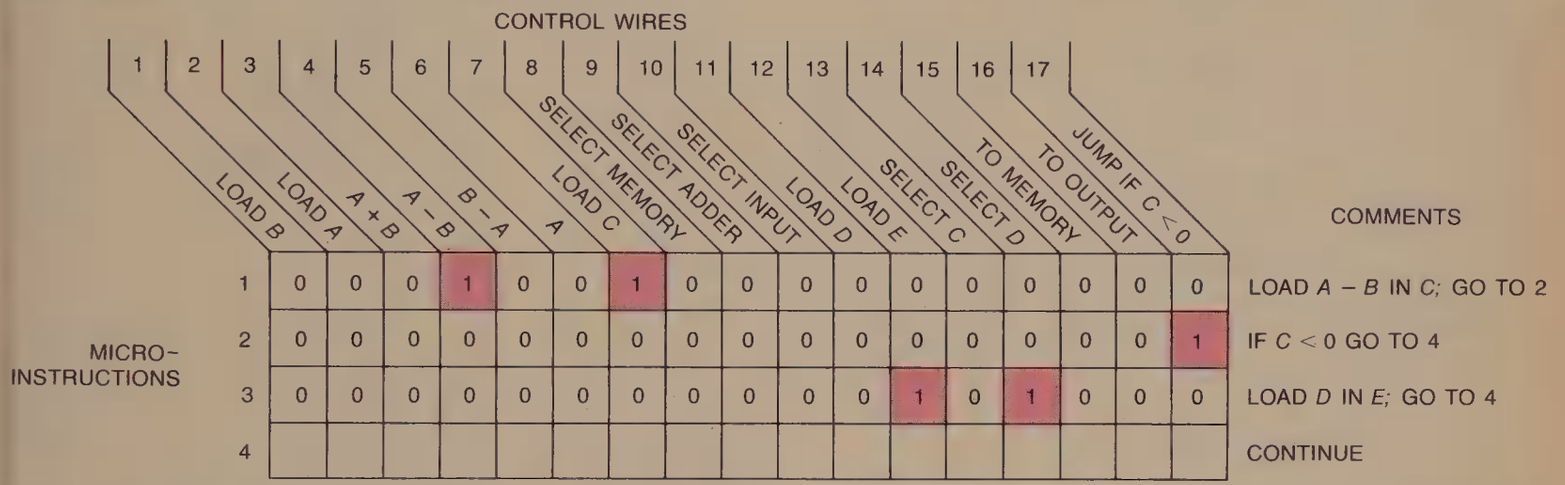
The hardware equivalent of the control matrix is a simple, repetitive memory structure. The content of each cell in a row determines the state of the corresponding control line for the duration of one clock cycle. The pattern of digits that constitutes a macroinstruction serves merely to select the appropriate row (or sequence of rows) in the control memory. In other words, the macroinstruction becomes an address designating the row. Designing a control system then becomes more a matter of software than of hardware; the complexity lies in choosing the right pattern of 1's and 0's to store in the memory rather than in specifying the right combination of hard-wired circuits to generate the control signals.

Wilkes recognized the similarities between his approach and conventional programming, and so he borrowed terms from programming to describe his ideas, in each case adding the prefix "micro-" to indicate the elementary nature of the control function. Hence "microprogramming" and its family of related terms. In particular Wilkes named each row of squares in the control matrix a microinstruction, and each sequence of rows that executes a single macroinstruction he called a microprogram. The memory structure designed to hold the microprogram is called the microprogram memory. The concept of microprogramming made it easier to understand the control function and, by replacing the complex control circuitry with a repetitive array of memory cells, made it easier to build the hardware. Most important, it gave the computer a new flexibility: the control system could now be changed without redesigning the hardware.

Consider a sample microprogram for the simplified processor discussed above [*see top illustration on opposite page*]. Suppose the operation controlled by the microprogram is to subtract the number stored in memory register *B* from the one in the accumulator register *A* and then to store the result in *A*. By examining the data path in the hardware diagram one can see that the device labeled "adder" is capable not only of



SAMPLE MICROPROGRAM for a single macroinstruction that can be executed by the simplified processor shown in the top illustration on the opposite page consists of a sequence of rows in a control matrix, or rectangular grid, in which each row corresponds to a microinstruction and each column is associated with a control wire. For each control wire that must be on during a particular clock cycle a 1 is written in the corresponding square and for each control wire that must be off a 0 is written. (Here "on" is also signified by color and "off" by white.) The operation controlled by the microprogram is to subtract the number stored in memory register *B* from the one in the accumulator register *A* and then store the result in *A*. As in the case of macroinstructions, the comments at the right are just reminders.



DECISION-MAKING MICROPROGRAM utilizes the augmented control hardware shown in the bottom illustration on the opposite page. The operation controlled by the microprogram is to load the number in register *D* into register *E* if *A* is greater than or equal to *B*. The conditional jump calls for an additional square to give the command to jump to microinstruction *4* if the number in register *C* is negative (column 17). Thus the conditional-jump macroinstruction comes down to a choice between microinstruction *3* and microinstruction *4*.

adding the two numbers but also of subtracting either one from the other. Before the result can be put back in *A*, however, it must be held temporarily in register *C;* hence register *C* must be loaded and then unloaded into the data path that leads back to *A*. The subtraction operation calls for two microinstructions. The first microinstruction tells the adder to subtract the number in *B* from the one in *A* and to put the result into *C;* accordingly control wires 4 ("*A* − *B*") and 7 ("LOAD *C*") are turned on. The next microinstruction moves the information from *C* into *A* by opening a gate in the wire from *C* in order to pass the information along to *A;* hence control wires 13 ("SELECT *C*") and 2 ("LOAD *A*") are turned on.

Now consider a somewhat more complicated microprogram for the same processor. The second microprogram calls for a decision-making process similar to the conditional jump in the income-tax program. The decision comes down to a choice between two microinstructions. For this purpose one more square must be added to the microinstructions and one more control wire must be added to the processor. The augmented control hardware is shown in the bottom illustration on the opposite page and the revised microprogram in the bottom illustration on this page.

The operation controlled by the second microprogram is to load the number in register *D* into register *E* if *A* is greater than or equal to *B*. First *B* is subtracted from *A* and the result is stored in *C*. If the number in *C* is negative, it follows that *A* is less than *B* and so the decision is to jump over the next microinstruction and resume the microprogram at microinstruction 4. If the result of the subtraction is a positive number or zero, on the other hand, *A* is greater than or equal to *B*, and the microprogram proceeds to the next sequential microinstruction, which loads *E* from *D*. A square is added to the microinstructions to give the command to jump to microinstruction 4 if *C* is negative, and a corresponding control wire is added to the processor's hardware.

A macroinstruction is said to be "interpreted" by the microcode to yield the control signals that actually manipulate the flow of information in the data path. Thus the sequence of microinstructions in the second example above is the microcode interpretation of the conditional-jump macroinstruction written as "JUMPIF" at step 19 of the income-tax program. The microprogram deals with the simple arithmetic test presented by the programmer by directing the hardware to test the number in the *C* register to see if it is negative, and if it is, to execute the appropriate conditional-jump microinstruction. For the hardware to perform such a test it must feed information from the data path back into the control system. The hardware needed to determine whether the number in *C* is negative is simply a wire from *C* that represents the sign of the number in that register: the wire is off (0) when the sign is positive (implying that *C* is greater than or equal to zero), and it is on (1) when the sign is negative (implying that *C* is less than zero).

The substitution of simple, repetitive memory structures for complex hard-wired control circuits yields two main advantages: it makes it easier to understand and build the control sys-

tem, and it makes it easier to modify the system. One can correct a mistake in a microprogrammed control system simply by changing the contents of the memory. For example, suppose one decides to change the preceding microprogram from one that loads E from D if A is less than B to one that loads E from D if B is less than A. The first microinstruction in the original microprogram loads C with the result of $A - B$. If one turns off control wire 4 and turns on control wire 5, however, then the microinstruction would load C with the result of $B - A$. In either case the next microinstruction determines whether the number in C is less than zero and jumps if it is. Hence one can correct the operation by changing only two control wires in one microinstruction. With a nonmicroprogrammed control system one would have to change the structure of the hardware to make the same correction.

Besides reducing the cost of correcting the control system the introduction of microprogramming makes it possible to change the control system completely by changing the total contents of the microprogram memory. Thus a single hardware system can be made to serve many different functions.

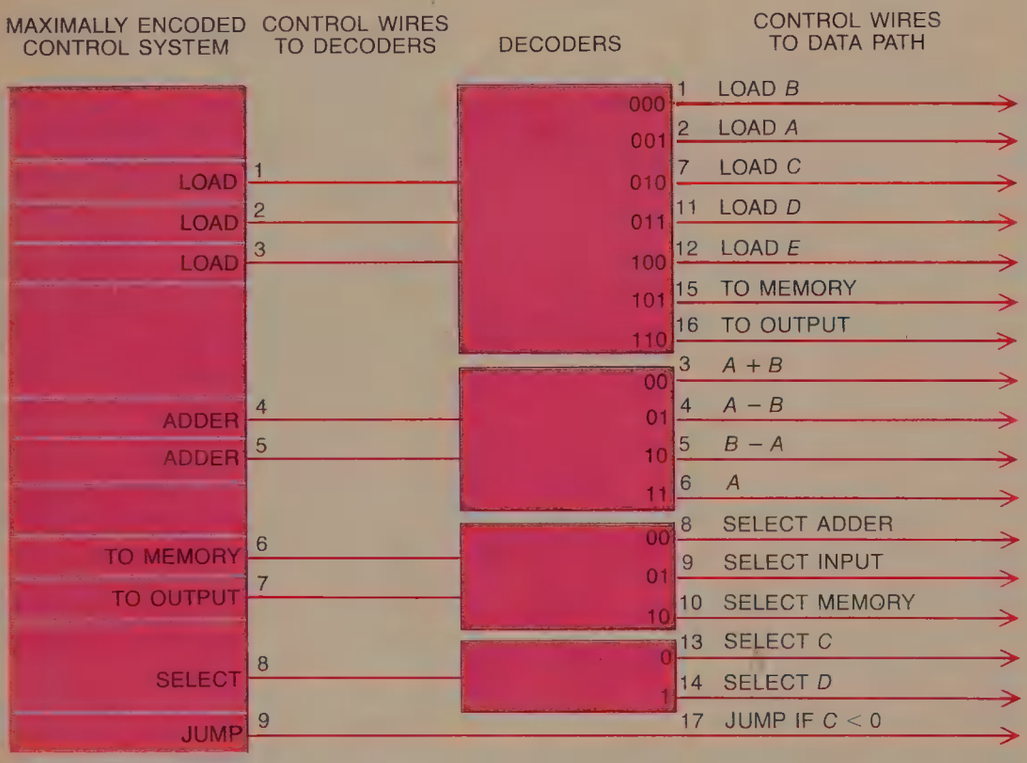
The advantages of microcode were apparent almost from the beginning. Wilkes's idea, however, was somewhat ahead of the hardware technology. Microprogramming calls for fast and inexpensive memory structures, and they were not available in the 1950's. Furthermore, it was generally thought at the time that any computer designed to take advantage of this ingenious approach would be slower than a machine with a hard-wired control system. As a result microprogramming remained little more than an intellectual curiosity for more than a decade.

It was not until the early 1960's that microprogramming emerged into the limelight, largely as the outcome of a major business decision by the International Business Machines Corporation. In the late 1950's IBM had in effect become many small companies, each one marketing its own type of computer. The computers differed not only in size, speed and price but also in the set of macroinstructions provided to operate them. Because of the differences in macroinstruction sets, programs devised for one type of IBM computer could not be run on another IBM computer. Obviously it would have been much easier to sell computers that operated with standardized software and that differed only in size, speed and price. The problem was to convince the managers of the subsidiary "companies" that it was possible for both the designers of the fastest computers and the designers of the cheapest computers to build efficient machines that relied on the same set of macroinstructions.

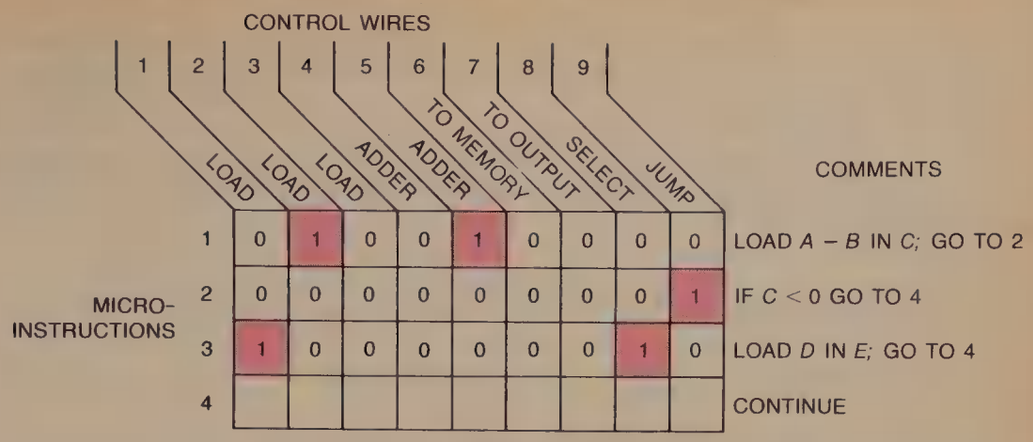
As it happened, the managers of the subsidiaries agreed to the innovative concept of a family of software-compatible computers only because they were persuaded that microprogramming made such a plan feasible. If one could microprogram the same hardware to execute different sets of macroinstructions, it was reasoned, then one should be able to microprogram different computers to execute the same macroinstruction set. The perceived marketing advantages of the family concept were so great that IBM made a major investment in the development of memory technologies that could lead to fast and inexpensive microprogram memories. Finally, in 1964 IBM announced a family of seven software-compatible computer models that varied in speed by a factor of 300 and in cost by a factor of 100.

Almost as an afterthought it was decided at IBM that microprogramming could help with the difficult task of making the programs for the old computers compatible with the new ones. Accordingly several models in the new IBM family ended up with two microprograms: one for the new, improved set of macroinstructions and another for the old macroinstruction set. This approach enabled customers to run their old programs without modification on the new hardware. The method proved so popular that for the first few years it was difficult to say whether the new IBM machines were spending more time running the microprogram for the new macroinstructions or the microprogram that emulated the old ones.

Most major computer manufacturers soon followed the lead of IBM in relying on microcode to produce a family of software-compatible machines that vary over a wide range of speeds and prices. The main difficulty encountered along the way was in duplicating the expensive memory technology invented by the engineers at IBM. Rapid progress in microelectronic technology eventually solved the problem. Every year for more than a decade the semiconductor industry was able to double the amount of information that could be stored on a single memory chip. This technology solidified the position of microprogramming as the control method of choice through the 1970's. The exceptions were "supercomputers," the fastest and most expensive processors. In that case special-purpose control is still



NUMBER OF CONTROL WIRES can be reduced by further encoding the control signals and introducing suitable decoding devices. In general this approach involves finding two wires that are rarely both ■■ for the same microinstruction and eliminating one of them, making the other serve a dual function. This schematic hardware diagram depicts a maximally encoded version of the processor shown in the two illustrations on page 54. The control wires to the data path have been rearranged to emphasize the wires that have been encoded together in the new control system. The additional encoding generally results in ■ control system that is slower than ■ minimally encoded one (in other words, one in which each microprogram square is matched by one control wire), because fewer operations can proceed simultaneously and because the control signals must pass through the decoders. In addition the resulting microprogram tends to have a larger number of microinstructions. Because each microinstruction has fewer binary digits, however, the cost of microprogram memory can be significantly reduced.



MAXIMALLY ENCODED MICROPROGRAM for the control system on the opposite page is said to be "vertical," because it usually has a larger number of shorter microinstructions, thereby making it taller and narrower than the corresponding "horizontal" microprogram. In this case the encoding was done in such a way as to require no additional microinstructions.

preferred for its greater speed, even though it is harder to build.

When microprocessors began to appear on the scene a decade or so ago, the question naturally arose of how to design their control systems. Repeating history, the first microprocessors were built with nonmicroprogrammed control systems. As the technology has matured, however, the advantages of microprogramming have persuaded most microcomputer designers to choose microcode.

In the meantime one kind of microprogramming has evolved away from the simple matrix concept in which each square is matched to an individual control wire. The improvements have come mainly from the effort to reduce the cost of the microprogram memory. In the execution of any single microinstruction it is generally the case that only a few components of the hardware are active; the control wires leading to all the other elements are off. As a result a microprogram tends to have islands of 1's scattered in a sea of 0's. If the same end could be achieved with less memory (that is, with fewer 0's), the result would be a less expensive memory. The solution is to supply the same information to the data path by further encoding the control signals. In general the additional encoding step involves finding two wires that are rarely both on for the same microinstruction and eliminating one of them, making the other serve a dual function. This approach tends to result in a larger number of microinstructions, but each microinstruction has fewer binary digits. A maximally encoded set of microinstructions is said to be "vertical," because the resulting microprograms are usually tall and narrow. A minimally encoded set of microinstructions is said to be "horizontal," because the resulting microprograms are usually short and wide. Minimally encoded microinstructions generally lead to faster computers, since more operations can be done simultaneously.

The growing importance of microprogramming has also stimulated interest in finding ways to enable computer users to write their own microcode. The potential advantages of transforming the computer from a general-purpose problem solver to one specifically tailored to a given application has made user microprogramming an attractive option. So far the best results obtained by users who have written their programs in microcode have shown an improvement in speed by as much as a factor of 10. In practice the gain in speed depends on the particular data path of the microprogrammed computer and on the application; in some cases the speed advantage is much less, and occasionally it may even disappear.

In most microprogrammed computers the microinstructions are stored in a form of read-only memory, whose contents are determined when the memory array is manufactured. Another kind of microprogrammed computer is capable of changing the contents of its microprogram memory in as little as a few thousandths of a second, while the computer is in operation. Such a machine is said to have a writable control store. The key to this approach is to have many different microcoded applications available to be loaded only when necessary. Endowing a computer with a writable control store is a way of removing the hardware barrier from user microprogramming.

In spite of these efforts microprogramming has remained so difficult that few users have taken full advantage of its potential. Therefore some investigators have attacked the problem of making it easier to write microcode. In many ways the situation is similar to one that arose earlier in the history of computer programming. Then the problem was the difficulty of arranging thousands of macroinstructions into a program that would operate without fail. This problem inspired programmers to write special programs to help write the application programs. The special programs translate high-level computer languages into the simpler macroinstructions a computer can execute. The high-level languages such as BASIC and Pascal are easier to understand and use, thereby making the task of programming less difficult, albeit at some sacrifice in the speed of the program.

For the past few years workers in microprogramming have been trying to create analogous translation programs for microcode. The problem is harder in this case because there are many more potential microinstructions than there are ordinary macroinstructions. A microprogrammed processor might have 50 control wires; since each wire can have one of two states (on or off), the total number of possible microinstructions would be 2^{50}, or approximately 10^{15}. Moreover, since the speed of the microprogram determines the speed of the computer, the demand for efficiency is much higher. The result has been the invention of high-level microprogramming languages aimed at balancing ease of microprogramming with efficiency. Recent developments have been encouraging, but general acceptance among the practitoners of microprogramming is still several years away.

Meanwhile investigators at the IBM Thomas J. Watson Research Center and my colleagues and I in the department of electrical engineering and computer science of the University of California at Berkeley have taken a different approach. We start by discarding the conventional macroinstruction set. We then write new programs that translate the standard high-level programming languages directly into microcode. Machines based on this principle are known generically as reduced-instruction-set computers, or RISC's. To test our ideas our group at Berkeley has designed our own microprocessor, which we call RISC I. In addition we have simplified the translation step by maximally encoding the microinstructions to achieve a highly vertical type of microprogram. In effect, all programs then become microprograms.

Recent tests by our group and by the workers at IBM have shown that a reduced-instruction-set computer would be from three to five times as fast as a conventional complex-instruction-set computer for many applications. Workers at Stanford University and at several industrial laboratories have begun projects to investigate this approach. If the preliminary findings are substantiated, a simpler type of computer might emerge, representing an unexpected advance in the continuing exploitation of the principle of microprogramming.

The Script of the Indus Valley Civilization

One of the four earliest civilizations has a script that has long resisted decipherment. A start on deciphering it is now made on the basis of facts and inferences from the archaeological record

by Walter A. Fairservis, Jr.

The oldest civilizations are those of Sumer, Egypt, China and the Indus Valley in the northwestern part of the subcontinent now occupied by India, Pakistan and Bangladesh. The writings of the first three civilizations can be read; the inscriptions of the fourth one remain largely enigmatic. As a result what is known of the Indus Valley civilization comes solely from the material objects brought to light by archaeology. It is ironic that in a part of the world noted for the antiquity of its literature even the most ancient accounts contain no valid reference to the first great culture that thrived there.

The culture of the Indus Valley civilization is called Harappan, after one of its two great cities: Mohenjo-Daro and Harappa. The difficulties that face those who would decipher Harappan writing seem at first virtually insurmountable. They derive in large part from the Harappans' limited use of their script. Their "texts" consist almost exclusively of brief inscriptions on seals and equally limited graffiti on pottery. No known inscription consists of more than 21 signs and the average text numbers only five or six. This, coupled with the fact that many of the signs are pictographic, has let the imagination of more than a few scholars run riot. Earnest attempts have been made to relate Harappan to Minoan, Canaanite, Hittite and even to the peculiar "writing" of Easter Island. Studies in recent years building on what the archaeological record reveals about Harappan life have now yielded a more rewarding approach to the decipherment problem. It is my intent to outline some of these achievements here.

In the decades since the two great Harappan cities were discovered in the 1920's, and particularly since the end of World War II, archaeologists in India and Pakistan have located nearly 1,000 other Harappan sites. They are spread in a wide arc from western India in the vicinity of the Narmada River northward across Gujarat and Kutch, through the Pakistani regions of Sind and the western Punjab and on into Indian Rajasthan and the eastern Punjab up to the vicinity of Delhi. Other settlements have been found along the coast of the Arabian Sea almost as far as the Iranian border and in Baluchistan; one settlement was even discovered close to the Oxus River, deep in central Asia, by a recent French archaeological mission.

Most Harappan sites are small, covering between two and five acres, and are near rivers or streams. There appear to have been three phases of settlement. Sites of the early phase are in or near the borderlands between the subcontinent and Iran. Those of the mature phase are more widespread. Most are in the valley of the Indus River itself, but such far-flung sites as those in Baluchistan and the one near the Oxus also belong to the mature phase. Those of the later phase tend to be far to the south and east of the Indus. Findings at these later sites also contain evidence that the older Harappan culture was in the process of merging with local pastoral and agricultural peoples. Such findings suggest Harappan civilization did not come to a sudden end but played a role in the development of the style of village life that is so characteristic of the subcontinent today.

What does the archaeological record tell us about Harappan life? First, it indicates that its mainstay was farming. Both grains and garden vegetables were grown, and cattle husbandry was central to the farm economy. Second, it indicates that the material culture of the Harappans was simple but not lacking in rich goods. They used copper and bronze for some tools and weapons and occasionally worked gold and silver into the beads that were their principal form of jewelry. At the same time the bulk of their artifacts consist of wood and bone, shell, flint and clay, all materials that were locally abundant. Among the beads, however, were some skillfully fashioned out of rarer semiprecious stones such as agate, carnelian and lapis. (The site near the Oxus was also close to early lapis mines.)

A further striking aspect of Harappan life was the extent of standardization. In architecture, building bricks were standard in size and were laid in standard ways. Drainage and sewer systems were standard in pattern. Dwellings were standard in dimensions and special structures (possibly public) were positioned with respect to private ones according to standard plans. Other aspects of the same phenomenon included standard weights and measures, pottery that was standard in shape and ornamentation and standardized artifacts such as ladles, loom weights and even toy carts. At the same time certain aspects of Harappan life suggest the later culture of India, for example the use of distinctive headdresses and of multiple bangles and necklaces and even the style of Harappan figurines.

So much for a broad summary of what is known. What is not known may be of equal significance. There is no evidence of rivalry between different Harappan states, of warfare, of major international trade or of the kings and courts and great temple complexes so characteristic of the other ancient civilizations of the Old World. The archaeological evidence reveals next to nothing about Harappan religion or political and social organization. Yet Harappan civilization was important to the civilizations that succeeded it. The Harappans cultivated cotton and perhaps rice, domesticated the chicken and may have invented the game of chess and one of the two great early sources of nonmuscle power: the windmill. (The other was the water wheel.)

When did this enigmatic civilization flourish? Although there is some controversy on the point, the mature Harappan phase seems to have extended from

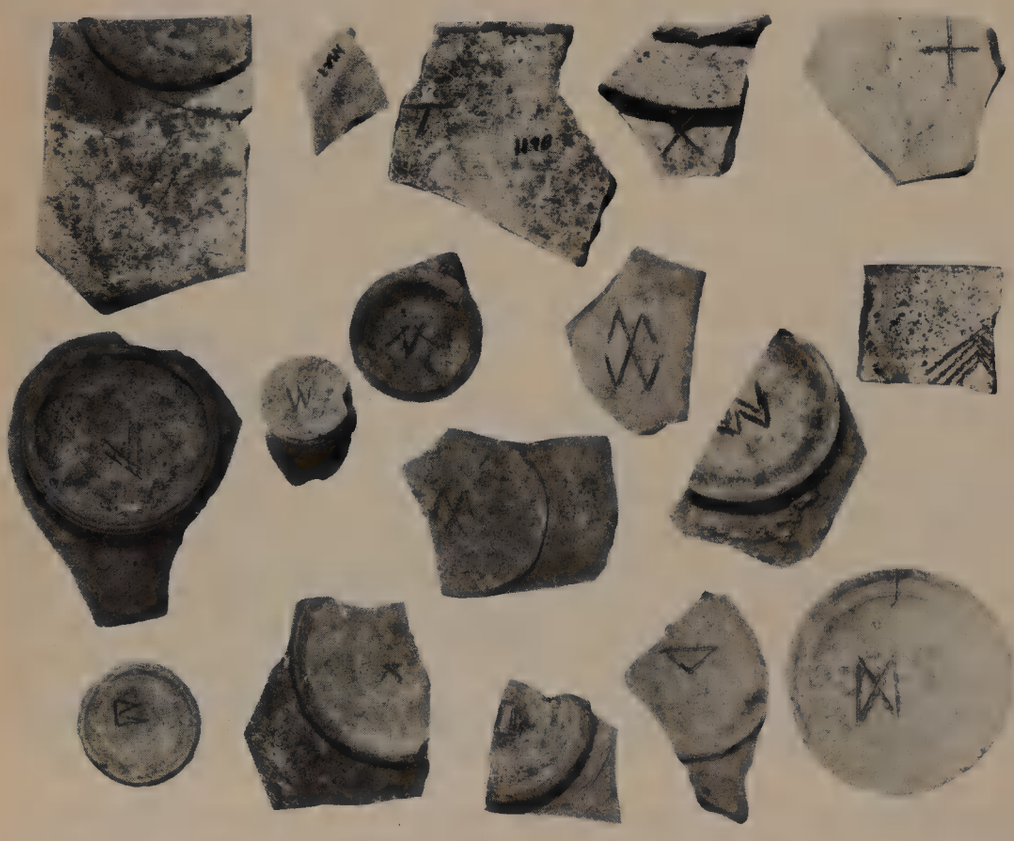
about 2200 to 1700 B.C. Most Harappan sites appear to have been occupied for no more than 200 years; they give the impression of a short-lived development characterized by substantial group organization and regular interaction between contemporaneous settlements.

The task of "reading" the Harappan inscriptions is sufficiently difficult to make many scholars believe it is impossible. The first difficulty arises from the fact that the inscriptions represent the unknown writing of an unknown language for which there are no bilingual texts, such as the Rosetta stone of Egypt or the Behistun monument of Iran. A second difficulty is the absence of long texts. The entire corpus of Harappan writing consists of some 4,000 seals, seal impressions (that is, impressions of seals) and graffiti on pottery; moreover, many of the seals are damaged and many of the pottery graffiti are interrupted by breakage. Still a third difficulty arises from the fact that Harappan civilization was not only geographically remote from the other civilizations of its time but also historically remote from later cultural developments on the subcontinent.

Fortunately the Harappan seals are inscribed with both writing and pictures, and the pictorial motifs can yield clues to what the writing means. Most Harappan seals are square or rectangular pieces of soapstone, a material that is easy to carve. On the back of most of the seals is a raised boss pierced with a hole for a carrying string; on the front is a combination of a negative picture and an inscription. The carvings presumably

SEAL OF THE HARAPPAN CULTURE in the archaeological collection of the National Museum of Pakistan in Karachi displays a zebu bull. The four signs across the top of the seal are among those that appear on the grid on pages 64 and 65. They are, from left to right, the triple stroke of the numeral 3, which also has the meaning "foremost"; a pipal-leaf figure and an arrow, with the meaning "mother" or "mistress"; a stick figure with "horns," indicative of a high-ranking personage, and a pot with handles, an honorific suffix with the meaning "high" or "superior." The seal evidently belonged to a high-ranking member of Harappan society, possibly a woman.

"LORD OF THE BEASTS," a famous seal in the archaeological collection of the National Museum of India in New Delhi, shows a human figure seated cross-legged and wearing a water-buffalo headdress. The "beasts" associated with the figure are a rhinoceros and a water buffalo (*left*), an elephant and a tiger (*right*) and a goat with its head turned (*bottom*). The goat may originally have been one of two, the second being lost when the seal was broken. The inscription can be rendered as "The Black One, the Black Buffalo *an-i*(*l*), the High One, the Lord of Chiefs."



GRAFFITI ON POTTERY, such as those on these sherds found in Baluchistan, are a second source of Harappan signs. They are not as numerous as the inscriptions on seals and seal impressions, but they indicate that the inscriptions were meant to be read from right to left.

identified the owner of the seal, so that when the seal was pressed into soft clay, the imprinted object was recognized as somehow associated with the owner.

A majority of Harappan seals display one or the other of two distinct pictorial motifs. The first involves an animal. It is usually a long-horned bull, but humped zebus are also portrayed, as are water buffaloes, goats, short-horned bulls, rhinoceroses, tigers, gavials (the river crocodiles of India) and elephants. Whatever the animal, an object was shown in place before it. The object shown with wild or dangerous animals is platterlike. The object shown with the domestic animals is either a basket or (particularly in front of the long-horned bull) a "stem" emblem such as one also shown being carried in processions.

A few of the animal seals show their subjects in groups. One well-known example is centered on an anthropomorphic figure, sometimes called the "Lord of the Beasts," seated cross-legged and wearing a water-buffalo headdress. To the right of the figure are an elephant and a tiger; to the left are a rhinoceros and a water buffalo. The seal is damaged but below and to the left of the central figure's feet can be seen a goat looking over its shoulder and to the right are what may be the horns of a matching goat figure, otherwise obliterated. A less elaborate seal centers on a pipal, or sacred fig tree (*Ficus religiosa*); the heads of two long-horned bulls are shown growing up from its trunk.

Another "animal group" depiction appears on one prismatic seal. (A few Harappan seals were made in the shape of a prism and a few others were cylindrical.) In a row on one of its three faces are an elephant, a rhinoceros and a tiger; a fourth figure in the row is too worn to identify. Presumably this is the "wild animal" face of the seal. On the second face is another four-animal procession, none of them apparently "wild." On both faces, however, a fifth animal is depicted above the animal procession. It is a gavial, self-evidently a member of the "wild" group. On the "wild animal" face the gavial is shown with a recognizable fish in front of its snout and what may be a fish behind it. On the second face no recognizable fish appears.

A fourth seal, even more elaborate than the "Lord of the Beasts," introduces a "worshiper" element. It repeats the depiction of a pipal tree, this time at its upper right corner. Between the branches of the tree stands a horned anthropomorphic figure. Facing the horned figure is a kneeling one, skirted and thus presumably female; to the left of the kneeling figure is a large goat. Seven skirted figures occupy the bottom half of the seal, their hair dressed in some kind of long "ponytail." A rather gruesome depiction on a fifth seal shows several similarly coiffed figures, one of

them wearing a skirt, being attacked by a water buffalo.

These motifs suggest something about Harappan social organization: that individual seal-bearers belonged to groups that transcended normal familial lines. For example, all individuals with a rhinoceros on their seal may have had some social tie in common. What was held in common could have been membership in some superfamilial group such as a clan or a club. The existence of procession scenes on a few seals, where animal effigies seem to be carried as standards, adds strength to this concept of superfamilial groups. The groups in turn may have been part of a larger two-part grouping, as is suggested by the motif of the "Lord of the Beasts" on the one hand and the motif of the "Pipal-deity worshipers" on the other. Such a possible structure is familiar to anthropologists as what is termed a moiety: a society characterized by the classification of clans or similar subsidiary organizations into two groups, "halves" that usually intermarry.

If one accepts the working hypothesis that the pictorial material on each seal identifies its bearer as to clan and moiety, it logically follows that the part of the seal devoted to Harappan script could be concerned with identifying the bearer as an individual. The script might, for example, give the individual's name, occupation, place of residence, rank or title and similar information. In support of such an interpretation the study of seal texts reveals considerable variation in the sequence of individual signs and yet a frequent repetition of certain of them. The hypothesis gives the would-be decipherer some basis on which to proceed. Just as Michael Ventris knew that at least some of the Linear B texts found on Crete were inventories of material objects, so the decipherer of Harappan script can assume that what to search for are proper names and their embellishments or other statements identifying the individual.

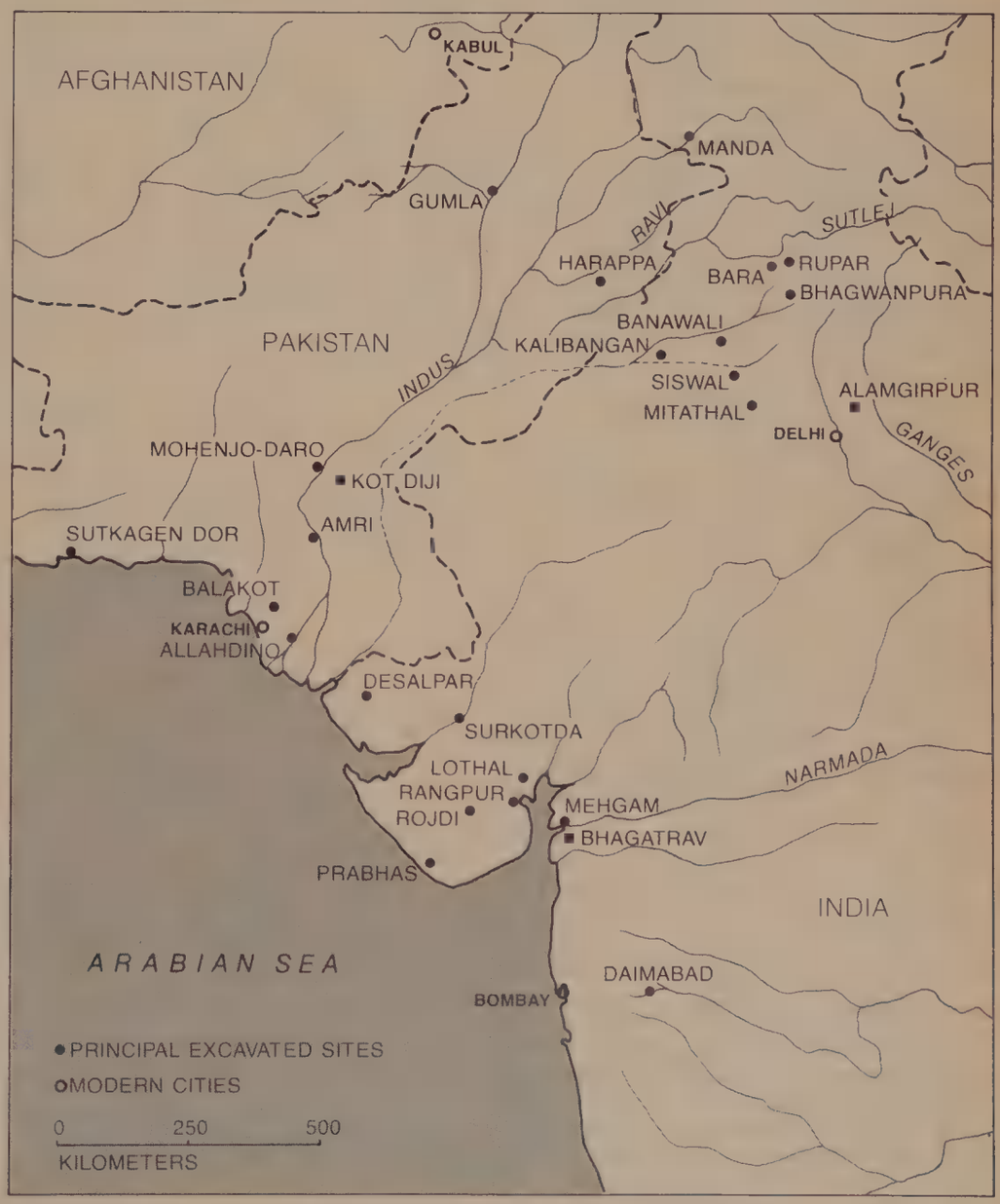
As long ago as the 1930's the British scholar G. R. Hunter had identified a total of 396 separate Harappan signs. More recent work has added another 23, so that the script is now known to incorporate a total of 419 signs. Statistical analysis shows that they occur 13,376 times in 2,290 known texts. Of the 419, 113 signs occur only once, 47 occur twice and 59 occur fewer than five times. In effect this means that the remaining 200 signs were in more or less general service, and analysis shows that fully half of them are combinations of the remaining half.

These findings demonstrate that Harappan writing was neither alphabetic, as Sanskrit is, nor logographic (that is, having one character for each word), as Chinese is. This places Harappan writing in the category known as logo-syllabic, meaning that some signs represent words and others serve purely for their syllabic values, or sounds. Other examples of this kind of writing are Egyptian hieroglyphs, early Sumerian ideographs and modern Japanese. The fact that half of the Harappan signs in common use were combinations of other commonly used ones suggests that the writers exploited such combinations to express ideas (as the Chinese do when they pair the characters for sun and moon to represent the word "brightness") and to combine syllabic sounds to "spell out" a word.

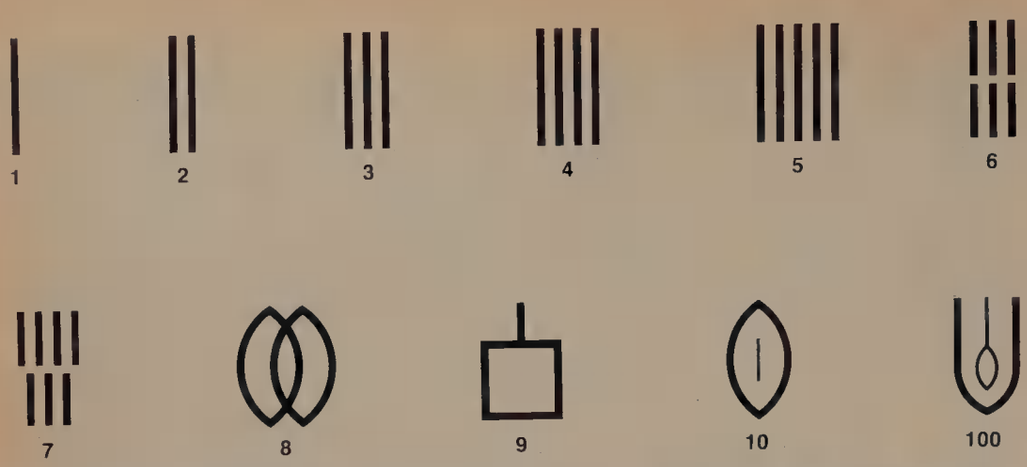
A crucial part of any writing system is the series of devices used to indicate gender, to distinguish between singular and plural, to establish the case of a verb and so on. The identification of these devices goes a long way toward establishing the relations between the graphemes, or individual components, of words and the language the graphemes represent. Now, the Harappan texts exhibit certain regularly paired signs, much as in the English reiteration of titles such as "His Majesty" and "Her Grace." It is also notable that certain signs appear in the middle of a text but rarely at the beginning or the end, whereas with other signs the reverse is true.

The usual ordering of signs and the identity of characteristic pairings may be established with a grid. Since the Harappan inscriptions are commonly brief, such a grid can consist of a relatively limited number of vertical columns. In fact, 14 columns were found to be sufficient. Of the first 17 inscriptions selected for horizontal grid display, most consisted of only five or six signs; thus they were entered near the center of the column array, allowing any longer lines to extend to the right or the left.

On entering the texts on the grid it became evident that certain signs appeared regularly in most of them. Con-



HARAPPAN SITES have been found from almost far south as Bombay to far north central Asia and from as far east Baluchistan to far west Delhi. Twenty-five of the principal excavated sites are shown on this map, including the great cities of the civilization: Mohenjo-Daro and Harappa. Most Harappan sites are small, covering between two and five acres.

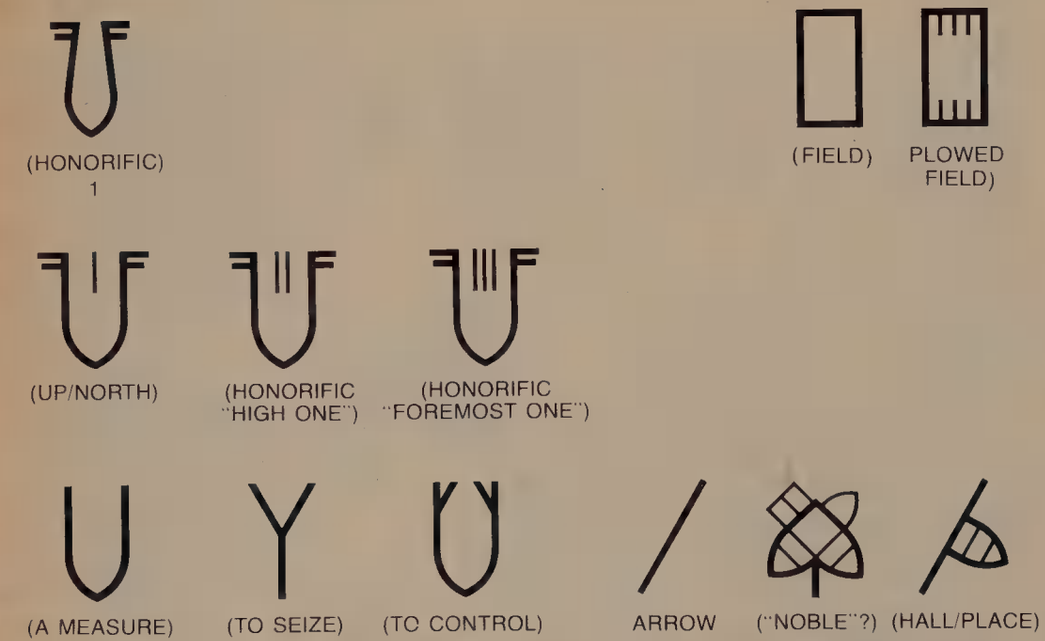


HARAPPAN NUMERALS, perhaps originally part of a system with a base of eight, are a plausible set of vertical strokes for 1 through 5 and a sextuplet of short strokes for 6. The sign for 7 is a similar set of short strokes. The signs for 8 and 9 are respectively pictographs of a double loom and a "foundation post." The sign for 10 is a single loom with a short stroke inside it. One other identified numeral, the sign for 100, is a pictograph of a mortar and pestle.

sider the "pot" signs in column 5, the "loom twist" signs in column 8 and the "two-stroke" signs in column 10. When two of these three signs appear in the same text (as in lines *b*, *e*, *h* and *p*), they are always in the same right-to-left order with respect to each other no matter what other signs are included in the inscription. Regularity of position evidently governed their relations. Accordingly in setting up the Harappan grid such signs were placed in the columns noted above even when only one of them appeared in a particular text.

What then became clear was that these signs were the most numerous among those in the 17 selected texts. The pot sign in column 5 appears there 10 times, and three variants appear in two other columns. The loom-twist sign in column 8, variants included, appears six times, and the two short vertical strokes of column 10, variants included, appear eight times. Certain other signs utilizing vertical strokes are not as fixed in their relative positions. Whereas the group of one or two short single strokes can be accommodated in column 10, those groups consisting of one, two, three or more long strokes fall outside the central columns of the grid.

By now, acquainted with the appearance of some 50 Harappan signs as they are displayed on the analytical grid, the reader may be wondering in which horizontal direction they are to be "read." On the basis of seal inscriptions alone this might have been a hard question to answer. Fortunately the graffiti inscribed on pottery supplied an answer. Studies by B. B. Lal and I. Mahadevan of the Archaeological Survey of India have shown that some graffiti have overlapping strokes. The overlaps demonstrate that the direction of writing was from right to left. That is why column 14 of the grid appears at its left margin and column 1 at its right margin. The reversal puts the reading order in the more familiar left-to-right pattern.



COMBINED SIGNS make up half of the total of the 200 signs most frequently encountered in the Harappan script. Four examples are given here. At the top left is one of the commonest script signs, a pictograph of a pot with handles (labeled with the Arabic numeral 1). Below it are three combined signs: the sign labeled *1* plus signs like those for the Harappan numerals 1, 2 and 3 but consisting of short strokes. Under each of these three signs appears its sense in English. Below the combined signs are seen, from left to right, the pictographs for a container and for a forked stick, and a third sign combining them. At the top right is the pictograph for a field, a simple rectangle. Beside it is a sign that combines the field sign with the strokes of the Harappan numeral 6 to form a new pictograph with the apparent sense of "plowed field." At the bottom right, from left to right, are pictographs for "arrow" and for "pipal-tree leaf." The third sign, combining "arrow" with a partial "pipal leaf," seems to indicate "hall" or "place."

We now come to the most complex part of the decipherment problem: What was the Harappan language? When Ventris identified the language of Linear B as Greek, he had overcome his greatest difficulty. The archaeological record of the region offers something of the same kind of help to those trying to pick a candidate for the unknown tongue. For example, the record shows that Harappan civilization was no sudden development. It had a long ancestry in the Indo-Iranian borderland, spread widely and eventually made its own contribution to the emergence of village India. This sequence suggests that the language spoken by the Harappans cannot have completely disappeared from the subcontinent.

Accepting such an assumption, which of the three principal families of languages spoken in the region might be related to Harappan? One candidate is Munda, a family of languages (spoken largely in eastern India) that seems affiliated with certain languages of Southeast Asia. Studies of the earliest forms of Munda, however, find little in its vocabulary that comports with what archaeology tells us about the culture of the Harappans. Another candidate is Indo-Aryan, a family of languages that traditionally came to India in the middle of the second millennium B.C. Its earliest literary expression, the Rig-Veda, however, describes a basically central-Asian culture quite different from the Harappan. The third candidate is Dravidian, a language now spoken mostly in southern and southeastern India but also still found in pockets in northern India and in Baluchistan.

A fourth possibility, of course, is that Harappan is related to none of the above. It is nonetheless worthwhile to consider the candidacy of Dravidian more closely than that of either Munda or Indo-Aryan. In addition to the northern forms (Kurukh, Malto and Brahui) about 25 Dravidian languages are still spoken. Indeed, the major families (Tamil, Malayalam, Kannada and Telegu) are spoken by more than 100 million people. Thanks again to the archaeological record it is not necessary to ac-

cept the candidacy of Dravidian blindly.

In 1974 excavations at the Harappan site of Allahdino, near Karachi, unearthed an ivory fragment with a half-circular cross section. There were holes on one side, apparently drilled to hold pegs; on the other side were two lengthwise parallel grooves. The object was similar to a large number of ivory rods or sticks excavated in the 1930's by one of the later workers at the great Harappan city of Mohenjo-Daro, E. J. H. Mackay. The Mohenjo-Daro ivories were not, as has been asserted, merely gaming pieces. For example, one of them, square in cross section, had a series of alternating circles and crescents cut into one side.

On some of the Mohenjo-Daro ivories is a sign seen in the grid (column 6, line *c*) and another sign resembling some form of plant. In various seal texts these two signs are associated with vertical strokes that range from one to seven in number and with five other signs. (Two appear on the grid at column 7, line *c*, and column 11, line *q*.) Like the second sign on the Mohenjo-Daro ivories, the sign at column 6, line *c*, appears to represent a plant, probably a stalk of grain. Considering the association of both signs in the seal inscriptions with what appear to be numbers and the marking of one of the ivories with circles and crescents that approximate in number a lunar (crescent) month of 30 solar (circle) days, the two plant signs might logically be taken to represent a word meaning both grain and month (or moon). Furthermore, the vertical strokes associated with both signs end after reaching a total of seven, suggesting that some other sign served for the number eight and possibly that the Harappan number system was to the base eight.

What language has a word for grain that also means month or moon and is associated with a base-eight numerical system? On the first point *nel* means rice in five Dravidian languages and *nilā* or *nela* means moon in three of the same five and in five others as well. It is also the word for month in some of the same languages. On the second point, a student of Dravidian language, Kamir V. Zvelebil, has pointed out that the original Dravidian number system was indeed probably to the base eight: the count to 10, used for conformity with the base-10 number system today, goes literally (in translation) "one," "two," "three," "four," "five," "six," "seven," "number," "many minus one" and "many."

Over the years scholars have noted in Dravidian a number of homophones: words with the same sounds but with different meanings. For example, the word for the shoulder pole from which pots are suspended is *kā*. It is also the word meaning guardian, or protector. The common word for fish is *mīn*, which is also the word for star. The word for 100 is *nūṟu*, which is also the word meaning to grind or to powder.

The homophonic, or rebus, principle is found in a number of ancient languages, including the hieroglyphs of Egypt. The key aspect of the homophonic principle is that it seeks a syllabic equivalent (say a picture of an eye to mean "I") rather than merely being a picture of something. For example, an early Egyptian ruler, Nr-mr, was represented in hieroglyphs by the sign for a catfish (*nr*) and the sign for a chisel (*mr*). The representation was not meant to suggest that he be called "the catfish-chisel one" but rather that his name sounded like "catfish-chisel."

Does this suggest a basis for deriving a Harappan syllabary? I shall put the point to trial in what follows, first warning that the acid test of any decipherment is its consistency. If a decision is made to declare a certain symbol equivalent to a selected sound or sense, it cannot later be shifted to represent some other sound or sense. Inner logic is basic to all writing systems; when decipherers arbitrarily alter the values of symbols in order to fit the model they prefer, their work is spoiled.

For a start consider the sign shown in column 7, line *h*. Pictographically it could be taken to represent a mortar and pestle. Linguistic reconstructions suggest that this utensil was given the syllabic value *nūṟu* in Dravidian. This, as we have seen, is also the sound of the verb to grind or powder and the numerical noun "100." Next consider the sign that figures so prominently in column 5 of the grid. Some years ago Mahadevan conjectured that this sign, which most

ADDITIONAL PICTOGRAPHS to be found on the grid on the next two pages include those illustrated here together with some combined forms. In the top row, from left to right, are "bow," "arrow," "man" and the combination "bowman." In the second row are "comb," a third human stick figure holding a comb (a combination that appears to indicate "scribe"), a fourth stick figure with "horns," a fifth stick figure with an elaborate coiffure ("woman") and a stick figure with a loaded shoulder pole (a combination appearing to indicate "guardian"). In the third row, from left to right, are "spear" and the shorthand for "spear," "point" (*color*), an honorific; a "loom twist" plus human arms, a second honorific pertaining to rulership, and the two signs combined. In the fourth row are two stalks of grain (both signs have the interchangeable sense of "grain" or "month"), a sign with the sense of "river" or "water" and a spiral with the sense of "surround" or "surrounding." In the bottom row are five astronomical pictographs. From left to right are seen the sun, the moon, sunrise or sunset, the crescent moon and a star.

frequently appears at the end of seal texts, was a pictographic representation of a pot with handles. He went on to point out that various Dravidian words for this kind of pot were homophones of words meaning male, including the common honorific suffix *aṉ* used with male personal names at least as long ago as early in the Christian Era. The fact that such a sign terminates seal texts hypothesized to contain personal names strengthens Mahadevan's conjecture.

Now, another terminal sign on the grid (column 5, line *f*) is a pictograph of a human figure carrying a shoulder pole with a pot at each end. As we have seen, the syllable for a pole with pots alone, *kā,* is a homophone of the Dravidian words meaning "to guard" or "to protect." Furthermore, a common Dravidian word for man is *āḷ.* The combination of these would allow the reconstruction of the sign at column 5, line *f,* as a two-syllable word, *kā-āḷ.* In the Dravidian languages, for the sake of euphony, such adjacent vowels are separated by a consonant, either a *v* or a *y.* Thus the restored word, *kāvāḷ,* could be translated as "one who guards or protects," a statement of personal identity suitable for entry on a private seal.

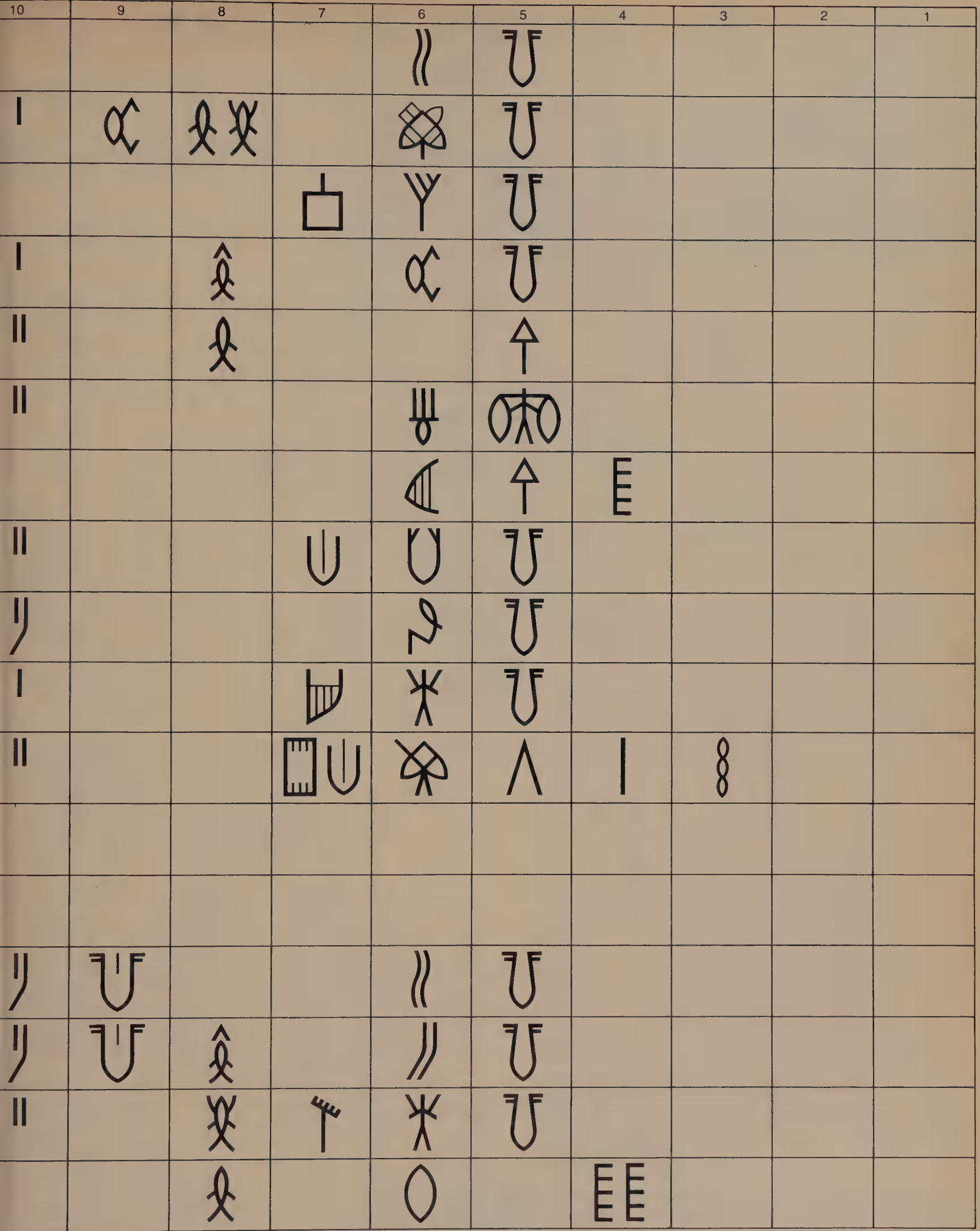
To give two further examples before turning to trial readings of some of the seal texts on the grid, two of the "number" signs—the sign in column 12, line *k,* presumably meaning three, and a second sign consisting of four vertical strokes, presumably meaning four—also have plausible homophones. The first homophone, *mu(n),* equates with the Dravidian word meaning "foremost" (*muṉ*); the second homophone, *nāl,* equates with the word meaning "good" (*nal*), both logical titular adjectives.

Let us now consider two short and two longer seal texts from the grid. The text designated Mohenjo-Daro 31 46 (line *g*) consists of three signs, putatively a musical instrument, a spear and a comb. The assigned Dravidian syllabic values are, in the same order, *pan, ār* and *ki(r),* in translation the word "sing," a third-person honorific and the word "mark," read as *Panār-ki(r).* The proposed translation is "Panar's seal," or literally "The singer's mark." Next, the text designated FEMD 590 (line *e*) consists of four signs. The first sign, a diamond shape, has the three-syllable value *āra-man,* the second, two short vertical lines, has the value *i(l),* the third, a "loom twist," has the three-syllable value *piri(key)* and the fourth is the same honorific *ār* as in the preceding text. This is read as *Āra-man-i(l) pirikeyār,* with the proposed translation "Belonging to the noble house (of the) Pirikeyar" ("pirs" and "pirikeys" are chiefs).

The third text, Harappa 72 (line *d*), consists of seven signs: an initial spiral (*cur*), an oval enclosing a short vertical

		14	13	12	11
a	FEMD 606				
b	HAR 16				
c	HAR 99				
d	HAR 72				
e	FEMD 590				
f	MD-31 121				
g	MD-31 46				
h	MD-31 26				
i	HAR 110				
j	HAR 102				
k	FEMD 111				
l	MD-31 69				
m	HAR 69				
n	MD 650				
o	MD 405				
p	FEMD 48				
q	ME-31 110				

ANALYTICAL GRID 14 columns wide contains 17 seal texts that range in length from two signs to nine. When "pot" signs were all assigned to column 5, "loom twist" signs and their variants to column [illegible] and the signs made up of two short vertical dashes to column 10, it became apparent that when two or more of these three signs appeared in [illegible] inscription, they always appeared in the same right-to-left order with respect to [illegible] another regardless of what other signs might appear in the text. Lines *b, e, h* and *p* provide four examples. The single short verti-



cal dashes in column 10 (lines *b, d* and *j*) reveal ▪▪ inflectional quality. The same sign also appears in combining form in column 12, line *d*, ▪▪ ▪▪ addition to ▪ "sun" sign; in column 11, line *k*, ▪▪ ▪▪ addition to ▪ "man" sign, and in column 9, lines *n* and *o*, as additions to "pot" signs. The sense of the single sign and its combination (lines *i, n* and *o*) appears to be that of ▪ possessive; the sense of the double-dash sign appears to be that of ▪ locative. Two of the texts, lines *l* and *m*, do not fall within the "normal" central grid distribution. Thus it can be assumed that they are not texts of the usual formulation. When two signs are found in one grid cell, the placement indicates that the signs are regularly found paired in seal texts. Readings of four of these 17 texts (lines *d, e, g* and *k*) are shown at the bottom of the next page.

HARAPPAN CONTACT WITH MESOPOTAMIA is evidenced by the discovery of Harappan seals in the region. The one shown here, with the seal at the left and the sealing impression at the right, was uncovered in 1975 in the ruins of a house of the Kassite period at Nippur, an ancient city in Iraq, by McGuire Gibson of the Oriental Institute of the University of Chicago.

dash (*pată*), a pipal leaf and an arrow combined (*ambara*), a short vertical dash alone (*ă*), a twist with a point above it (*āra-pirikey*), pincers combined with the sign for the sun (*paṭu-kāru*) and, as the terminal sign, the pot with handles (*aṉ*). This is a mouthful: *Cur patăm-bara-ă āra-pirikey patu-kāran.* The proposed translation is "Patukaran, powerful (noble) chief of the surrounding settlements." The fourth text, FEMD 111 (line *k*), is one of the longest of the translated seal inscriptions, consisting of nine signs. In the interest of brevity, only the syllables and their proposed translation will be given: *Munālă-i(l) nūṟu cāṟuvara amban ārōru malya,* or "Belonging to Munala, mistress of 100 plowed fields, noble first lady."

So far syllabic values have been assigned to nearly 100 seal and pottery signs and appropriate Dravidian homophones have been found for each. Translations such as those given here have been proposed for more than 100 seal texts. They range from such simple statements as "water-holder" (probably a proper name) to such sonorities as "Arasamban, High Chief (of) Chiefs of the Southwest, lineage of the Moon."

What such readings demonstrate about the Harappans is that a number of individuals (Arasamban among them) traced their lineage to such major figures of the cosmos as the Sun, the Moon and the Stars and perhaps also the Monsoon Rain. These may represent distinctions within each clan. Chiefs were associated with *aramani,* or chiefs' houses, which may have had both residential and administrative functions. In addition there was a high place (the "citadels" identified by archaeologists?) that served a special but so far undetermined function. An assembly area—an open court or a pillared hall (both known archaeologically)—suggests that an assembly of chiefs was a basic part of Harappan political organization.

Among other kinds of leaders known from the seals were heads of associations (guilds?) such as that of the coppersmiths, storehouse overseers, irrigation supervisors and landowners, a category that as we have seen included women. Religious references, however, are scarce. There is a possible "horned deity," referred to as "the copper one" or "the red one," and a possible "mother deity," but at the moment these distinctions are not confirmed. Lesser figures included drummers and singers, the drummers perhaps to summon assemblies and the singers perhaps to entertain or perform at ceremonies.

The seals attest further to a class of scribes, to people in charge of weights and measures and to supervisors of the distribution of stores, the grinding of flour and probably hunting operations. There were also captains of boats and custodians of fire. Many seals bear the sign *kā,* referring to guardianship, not in a military sense but more in terms of a responsibility for the care of crops and the preservation of herds and flocks. One of the gratifying near-proofs that the decipherments are on the right track comes from the prismatic seal where a gavial is twice seen above an array of "clan" animals. In Dravidian the word for crocodile is *mutalai.* The Dravidian word for first chief, *mutali,* is a close homophone.

Thus it appears that the language of Mohenjo-Daro and Harappa some 4,000 years ago was an early Dravidian tongue and that the Harappan scribes struggled to put that language into graphic form as a method of identifying the elite of the Indus Valley civilization. The Harappan civilization, however, was geographically so widespread that a number of non-Dravidian words must have entered the language, just as Dravidian words were later borrowed by the speakers of Indo-Aryan.

A great deal of additional decipherment remains to be done and no doubt what has been put forth here will be found to have flaws. What remains to be done is even more exciting. It leads toward a goal that until recently seemed impossible to attain: cohesive information on Harappan polity, social organization and ideology, and perhaps even stronger evidence that, as the parent of the village India of today, the Harappan culture never did disappear.

FOUR SEAL INSCRIPTIONS appear together with the syllabic value of each sign in Dravidian, a surviving family of early languages of the region. The shortest inscription (*a*), in the author's reading, is "The singer's mark." The next inscription (*b*), which repeats the third-person honorific "spear" sign, *ār,* is read "Belonging to the noble house (of the) Pirikeyar." The next inscription (*c*) is seven signs long and includes a variant of the "loom twist" sign of inscription *b.* It is read "Patukaran, powerful (noble) chief of the surrounding settlements." The last inscription (*d*) is nine signs long and includes the two-stroke sign *i(l)* of inscription *b* and an abbreviation of the "spear" signs in both *a* and *b* (*color*). It is read "Belonging to Munala, mistress of 100 plowed fields, noble first lady." More than 100 seal texts have now been read. The majority of the readings accord with the hypothesis that the seal signs identify individuals.

We discovered a new kind of brain cell.

You'll find it at places like the Institute for Advanced Study, where experts from a dozen different disciplines cross-pollinate their thoughts and ideas.

But as far as companies in America go, we at Kodak may be the only ones who can bring together our own experts in programming, optics, physics, organic and polymer chemistry, fibers, human engineering, imaging, marketing, paper technology ...we've probably skipped a few.

Call it synergism or whatever. But when these pros sit down at our table, they come up with some truly incredible results.

The Kodak Ektaprint 250 duplicator is a good case in point. It has grown out of more disciplines than any single Kodak product before it.

The nucleus is two 8085 microprocessors, with a combined yield of over 100,000 bytes, and responsibility for 134 separate functions.

There are all kinds of thoughtful touches. One is a programmable display that's as friendly as can be—with 94 different messages, not codes, to help anticipate and satisfy operator needs. And there's an internal diagnostic system that's amazingly specific.

Then there's single-pass duplexing. Lots of heads went together for this run-through-once, print-both-sides convenience. Collating and cover inserting, too.

A couple of other neat feats: Fiber specialists from our Eastman Chemicals Division in Tennessee helped create the "perfect" toner brush. In Rochester, our experts in 35 mm added sprockets to the belt drive. And a reflective "bathtub" was perfected by our processing group; it helps keep the paper path shorter and virtually trouble-free.

With all these new capabilities, you may think the 250 needs a genius to run it. Wrong. And that's the beauty. 99% of the duplicator is microprocessor-controlled. That means reliability, flexibility, productivity, easier operation. And time to think about more important matters.

For more information on how we plan to combine our talents for future applications in microelectronics, please send for our brochure ..."Resources To Innovate, Commitment To Serve..." (B3-190). Write to Eastman Kodak Company, Dept. GB-SA 6, 343 State Street, Rochester, New York 14650.

Kodak

Kodak. Where technology anticipates need.

© Eastman Kodak Company, 1983

SCIENCE AND THE CITIZEN

Under the Volcano

It has long been hypothesized that large volcanic eruptions can affect the climate worldwide for several years. By putting dust and ash into the atmosphere volcanoes are assumed to reduce the amount of sunlight reaching the earth's surface. The result could be a global cooling, although the mechanism has not been precisely determined. In part because of this imprecision the explosion last year of the Mexican volcano El Chichón has aroused considerable interest among geologists and meteorologists. Now Michael R. Rampino of the National Aeronautics and Space Administration's Goddard Institute for Space Studies and Stephen Self of Arizona State University suggest that the effect of El Chichón on global climate could be as great as that of much larger eruptions, which lowered mean annual temperatures in the Northern Hemisphere by about .4 degree Celsius.

El Chichón, which until the eruption was about 1,400 meters high, is in Chiapas, the southernmost state of Mexico. Chiapas is at the intersection of three major tectonic plates: the North American plate, the Caribbean plate and the Cocos plate, which lies under the Pacific Ocean off Central America. The subduction of the Cocos plate under Mexico is thought to be the cause of volcanic activity in a broad belt in southern Mexico; El Chichón lies at the southern end of the zone. It had been dormant throughout history until it erupted with great force on March 28 and again on April 3 and April 7.

Almost immediately after each of the major eruptions a large cloud of dust, ash and gas appeared in satellite photographs of Mexico. The three clouds, reaching a height of about 16 kilometers, soon merged into a single mass. The lower part of the combined cloud moved northeast toward the Gulf of Mexico; it was quickly thinned out by the settling of dust particles and the scavenging action of ordinary water clouds. The higher part of the cloud, extending into the stratosphere, was driven westward over the Pacific.

In mid-April the stratospheric mass was detected over Hawaii by measurements of the amount of solar radiation reaching the earth. In May the cloud was observed by workers in Ahmadabad, India, who reported that it extended from an altitude of 17 kilometers to 30 kilometers, a range that includes most of the stratosphere. It is thought the stratospheric cloud circled the earth in a narrow belt, then diffused to higher and lower latitudes and eventually covered the globe completely.

To what extent will the El Chichón cloud affect the weather? Writing in *Quaternary Research,* Rampino and Self have given a basis for answering the question by correlating the amount and the type of matter ejected in three big volcanic eruptions with subsequent changes in mean annual temperature in the Northern Hemisphere. Tambora, on Sumbawa Island in Indonesia, erupted in 1815 and caused a cooling of about .8 degree C. Krakatoa, in the Sunda Straits between Java and Sumatra, erupted in 1883 and reduced the average temperature by about .4 degree. Mount Agung, on Bali, erupted in 1963; again the temperature reduction was about .4 degree.

Most attempts to explain the effect of volcanoes on climate have focused on the total volume of matter ejected during the blast. Even though the Tambora, Krakatoa and Mount Agung eruptions had similar effects on climate, the volumes of rock ejected were very different: about 150 cubic kilometers for Tambora, 20 cubic kilometers for Krakatoa and one cubic kilometer for Mount Agung. In explaining the disproportionate effect of the eruption of Mount Agung, Rampino and Self cite the andesitic composition of the volcano. (Andesite is a volcanic rock intermediate in composition between the basaltic and the silica-rich types.) Andesitic rock tends to have more sulfur than the silica-rich rock of Tambora and Krakatoa. Analysis of rock samples has shown that the sulfur content of Mount Agung was 800 parts per million, compared with 380 for Tambora and 150 for Krakatoa.

In an eruption the sulfur is converted into sulfur dioxide gas, which reacts with water vapor in the stratosphere to yield a fine aerosol mist of sulfuric acid. Rampino and Self note that there seems to be a limit to how much acidic aerosol can be formed in the stratosphere; for this reason a large increase in the amount of ejected matter does not cause a corresponding increase in the amount of acid. Furthermore, above an altitude of 30 kilometers the droplets evaporate quickly; thus an exceptionally powerful eruption, which throws material high into the atmosphere, may yield a smaller quantity of aerosol than a less powerful one.

Rampino and Self estimated the amount of sulfur-bearing aerosol put into the atmosphere by each of the volcanoes by measuring the amount of sulfate deposited in the annual layers of the polar ice cap. The measurement showed that the aerosol yield from Tambora, Krakatoa and Mount Agung was in the ratio of 7.5 : 3 : 1, much smaller than the ratio of the total volume of ejected matter. Moreover, given the imprecision of the estimates, the sulfate ratio is close to the observed ratio of temperature reductions. Rampino and Self conclude that it is the acidic aerosol that makes the largest contribution to blocking solar radiation.

The eruption of El Chichón was small compared with the three historical examples: it is estimated to have ejected about .5 cubic kilometer of rock. Like Mount Agung, however, El Chichón is an andesitic volcano. Furthermore, it is "exceptionally rich in sulfur," even richer than the rock type would indicate, Rampino has said. "There may have been an input of sulfur from sulfates in rock below the volcano," he added. Samples taken near the volcano have suggested there is a stratum laden with calcium sulfate.

Because of El Chichón's sulfur content and because its eruption had the force needed to blanket the stratosphere, it may have an effect on climate as great as that of Mount Agung, Rampino said. The mean temperature in the Northern Hemisphere could be lowered by as much as .5 degree this year and next. It should be noted, however, that although such a cooling is measurable, it is of roughly the same magnitude as random year-to-year fluctuations.

Rapid Pulse

The study of pulsars, compact sources of periodic radio emission in the galaxy, has been disarrayed at least temporarily by the discovery of the fastest one yet: a pulsar in the constellation Vulpecula pulsing so rapidly that it may prove to represent a new class of astrophysical objects. When the newly discovered pulsar first came to attention, it was hidden in a region of radio emission designated 4C 21.53. Then the pulsar itself was identified and was designated 1937 + 214. Investigators who are trying to account for its extraordinary properties suspect it has had an eventful billion-year history. They call it a recycled pulsar.

What made 1937 + 214 stand out in the radio sky were two attributes of 4C 21.53. First, the radio spectrum of 4C 21.53 was known to decrease steeply with increasing frequency, a characteristic of pulsars. Second, 4C 21.53 appeared to twinkle at radio frequencies, much as the image of a star twinkles at optical frequencies. The optical twinkling is caused by irregularities in the atmosphere of the earth; the radio twinkling is caused by analogous irregularities in the solar wind. A source of radiation can twinkle only if it is smaller than a typical irregularity, and so the radio source in 4C 21.53 could have an angu-

Taste is all it takes to switch to Jim Beam.

KENTUCKY STRAIGHT BOURBON WHISKEY. 80 PROOF. DISTILLED AND BOTTLED BY JAMES B. BEAM DISTILLING CO., CLERMONT, BEAM, KY.

lar diameter no larger than one second of arc, or .0006 times the diameter of the sun. That too suggested a pulsar.

Yet no pulsar had been detected in 4C 21.53. Examining this puzzle in 1979, D. C. Backer of the University of California at Berkeley decided the reason might well be a phenomenon called pulse broadening. Here the radiation from a pulsar is disrupted by turbulence in the ionized gas that permeates interstellar space. The parts of a pulse that are not disrupted reach the earth with the least delay; the parts that follow zigzag trajectories take longer. The effect is to smear out the pulses, particularly if the pulses are short. Backer estimated that if a pulsar in 4C 21.53 emits pulses with a period of less than 10 milliseconds, it would not have been detected in the searches made up to then. This limit, however, was one-third the period of the fastest pulsar known.

In 1979 and again several times last year Backer and his colleagues undertook further searches. They examined radio wavelengths of about 20 centimeters, where the disruption caused by the interstellar medium is minimal. Last November a search succeeded. Backer and his colleagues found 1937 + 214 in signals detected by the giant radio telescope of the Arecibo Observatory in Puerto Rico. Its pulses have a period of 1.5578 milliseconds, and so it pulses 20 times faster than the fastest pulsar known before.

Pulsars are taken to be spinning neutron stars that broadcast electromagnetic radiation in a beam, like a lighthouse. Thus their pulse rate equals their spin rate. It follows that 1937 + 214 spins once every 1.5578 milliseconds, or 642 times per second. The rate is alarmingly fast. Suppose 1937 + 214 has the mass of the sun and a radius of 10 kilometers (values thought to be typical for a neutron star). Then at its equator the surface of the star is moving at 13 percent of the speed of light. Moreover, the rotational energy of the body amounts to 7×10^{51} ergs. "This energy," Backer and his colleagues write in *Nature,* "is comparable to the entire mechanical energy output of a supernova," the stellar explosion that it is thought can give rise to a neutron star. The rotation rate approaches a natural limit: if it were about three times as fast, centrifugal force would overcome the gravitational self-adhesion of the pulsar and it would fly apart.

All pulsars slow down as they lose energy, and in general the deceleration is greater for pulsars that are spinning faster. The rate at which 1937 + 214 is slowing, however, is the lowest ever observed. In November it was established that the pulse rate (and hence the spin rate) is decreasing by no more than 10^{-15} second per second. Even before a recent measurement by Backer and his colleagues, which indicates that the spin of 1937 + 214 is decreasing by only 10^{-19} second per second, it had become questionable whether the accepted hypothesis of how a pulsar is formed applies to the newly discovered one.

In the accepted hypothesis a supernova explosion leaves behind a stellar remnant (a neutron star) with a high rate of spin and an intense magnetic field. In effect the tiny neutron star preserves the parent star's angular momentum and magnetic flux. The neutron star loses energy in part because a spinning dipolar magnetic field (in essence a spinning bar magnet) can emit electromagnetic waves and in part because a spinning body that is not a perfect sphere is thought to give off gravity waves. When the neutron star is young and spinning fast, the rate of loss is great; therefore the rotation rate decreases and (in the view of some theorists) the magnetic field subsides.

How can 1937 + 214 be spinning fast and yet be losing energy slowly? In hypotheses advanced by several investigators, notably M. A. Alpar of Columbia University and his colleagues there and at Rutgers University, the answer is that the pulsar has been recycled. In one scheme 1937 + 214 was created in a binary star system when a supernova explosion left behind a neutron star. The companion star was undisturbed. The neutron star followed the normal evolution of a pulsar: it spun rapidly and lost energy, so that its spinning slowed and its magnetic field was dissipated. It may even have been extinguished as a source of radio pulses. Over billions of years, however, the companion star lost mass. Each increment of mass that passed from it to the neutron star brought with it an increment of angular momentum. The neutron star was thereby "spun up" again. Today its rotational energy is great but its magnetic field is small, so that it loses little energy. At some point the companion star was destroyed in a second supernova explosion, which left behind no solid remnant. By now even the gaseous debris has dispersed. Only the recycled pulsar remains.

Social Aging

On being cast out of Eden mankind was condemned to a life of toil and to the pain of childbearing. Ever since it has been the norm that almost all the years of adulthood are spent working and raising children. The pattern is now changing in the developed countries, but not everyone will welcome the change as a return to the Garden. Many people, after all, count work and children among life's satisfactions rather than among its burdens.

The causes of the change are demographic. The fraction of a population that survives to the threshold of old age has increased dramatically in the past 100 years; the life expectancy of an individual who has reached age 60 has also increased, although to a lesser extent. Furthermore, people are having fewer children, and the age at which they have their first child decreased through most of the 20th century; so has the age at retirement. As a result the period spent "at liberty," with neither parental responsibilities nor an obligation to work, has been greatly lengthened.

Massimo Livi-Bacci of the University of Florence discusses these trends in *Population and Development Review* and presents a quantitative analysis of their effects on the population of Italy from 1881 through 1981. The analysis is made somewhat uncertain by the limitations of the available data, which are taken mainly from census reports. For example, the Italian censuses have not recorded the age of parents when their last child is born, which determines the parents' age when all their children have left home. Nevertheless, the missing data can be plausibly estimated from the known age distribution of the population and the fertility rate.

Longevity has of course improved substantially. In 1881 the average life expectancy at age 35 was less than 30 years; in 1981 a man of 35 could expect to live another 37 years and a woman another 43 years. Livi-Bacci shows, however, that the number of years spent without dependent children or employment has increased by a much larger factor. "For men these years were an insignificant portion of life in 1881 (only one year), but they have gradually increased to 12 years in 1981, or one-third of [the expectation of life at age 35]. For females the increase has been extraordinary: from five years in 1881 to almost 23 in 1981, or more than half the expectation of life at age 35."

The result, Livi-Bacci notes, is a postponement of biological aging but an earlier onset of social aging, "the process of relinquishing meaningful social functions, with the ensuing increased risk, for each individual, of becoming prematurely obsolete, senescent and estranged from society. It is in the increasing number of years... liberated from the two fundamental functions of parental responsibilities and work that social aging makes its mark."

It seems unlikely that the workless and childless period at the end of a typical life will ever grow much longer than it is now. Indeed, natural demographic adjustments may shift some of the years of reduced responsibility to an earlier stage of life. "Youth is a very busy period in the life cycle," Livi-Bacci writes. "Education is a long process; leisure is attractive and varied; work is difficult to find but part-time jobs are easily and frequently taken and left. Thus responsible childbearing could be postponed to

THE SEDAN THAT DEFINED PERFORMANCE FOR A GENERATION OF CARS UNFAMILIAR WITH THE TERM.

"In the beginning," wrote Car and Driver, "when BMW said, Let there be sports sedans, there were many other companies that failed to grasp the importance of this edict."

What went ungrasped was the nature of performance itself. Which is not just road-holding prowess, or agility through corners, or acceleration, but all of these in tandem.

And that is also an excellent definition of the BMW 320i.

PERFORMANCE: THE UNABRIDGED VERSION.

Like all BMW's, the 320i isn't so much the sum of its parts as a perfect balance of them.

Its fuel-injected 4-cylinder engine is the basis for BMW racing engines.

Its gearbox is, as one critic observed, "the perfect complement to the (320i's) willing engine."

Which is complemented by a suspension that removes the treachery from even serpentine stretches of roadway.

The result, according to Car and Driver: "maybe the best-balanced small sedan around."

VALUE, ACCORDING TO THE 320i.

What is a car really worth? Whatever the market will pay. By that standard, the 320i is valuable indeed. According to the NADA Used-Car Guide, 320i's have traditionally enjoyed one of the highest resale values of any car in their price class.

It's also one of the best protected, with a 3-year/36,000-mile limited warranty and a 6-year limited warranty against rust perforation.*

All things considered, it's understandable why other car makers "are already boasting that their next generation of sporty coupes will feel and drive like their target car, the 320i" (Motor Trend).

If you're unwilling to wait that long for someone else's interpretation of a 320i, we invite you to test drive the original at your nearest BMW dealer.

THE ULTIMATE DRIVING MACHINE.

*Warranty applies to automobiles purchased from authorized U.S. BMW dealers only. See your BMW dealer for details. © 1982 of BMW North America, Inc. The BMW trademark and logo are registered. European Tourist Delivery can be arranged through your authorized U.S. BMW dealer.

Life is an open book. If only he'd open the book.

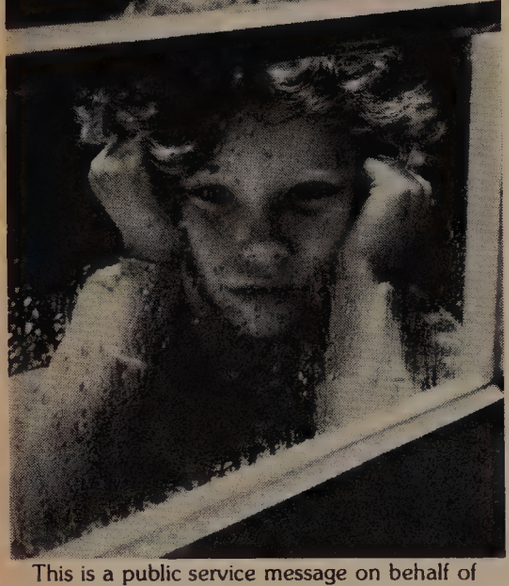
Reading is thinking. And learning. And growing. But a kid won't read if he doesn't want to. So we're giving kids the incentive. And the books.

We're RIF (Reading is Fundamental), a national program with hundreds of local, community projects that help kids help themselves to books. Books they can choose for themselves. And keep for their own.

RIF works. But RIF only works if people work, too.

That's why we need you—or your organization—to help start a RIF program in your community. We'll help you to start going and start growing. Write RIF, Box 23444, Washington, D.C. 20024

READING IS FUNDAMENTAL



This is a public service message on behalf of Reading Is Fundamental and this magazine.

a later stage in the life cycle (the trend has perhaps already been initiated)."

LEP and SLC

In the effort to accelerate particles of matter to ever higher energies one accelerator after another has been designed, built, exploited and then supplanted. The linear accelerator was followed by the circular design, which in turn was followed by the storage ring, where beams of particles collide head on. What may well be the last and biggest of the storage rings is about to be constructed. At the same time plans are under way to build a linear colliding-beam device, which may turn out to be the prototype of a new generation of accelerators.

The new circular machine is called LEP, or the Large Electron-Positron collider; it will be built by the European Organization for Nuclear Research (CERN) on a site straddling the French-Swiss border near Geneva. The storage ring will be in a tunnel 27 kilometers in circumference bored under the plain between Geneva and the Jura Mountains and under the mountains themselves. The particles in the ring will reach energies of about 50 billion electron volts (GeV) and will complete a circuit of the loop in less than a tenth of a millisecond. The other project is the Stanford Linear Collider (SLC), which may be added to the facilities of the Stanford Linear Accelerator Center (SLAC). It too would have a beam energy of about 50 GeV; LEP is scheduled for completion in 1987. The SLC will be completed in 1986, if it receives funding in fiscal year 1984.

The aim of a particle accelerator is to convert kinetic energy into mass and thereby create new forms of matter. Most physicists now classify the fundamental constituents of matter in three "generations" of the particles called quarks and leptons. All the members of the first two generations have been identified in experiments, but two particles in the third generation have not yet been seen: the quark called top and the lepton called the tau-type neutrino. They are among the particles that will be sought in the new colliding-beam laboratories. Also on the physicists' most-wanted list are certain particles that transmit forces between the quarks and leptons, notably the particle called the neutral weak boson, or Z^0.

In a modern linear accelerator electrically charged particles are injected into a long evacuated tube and are accelerated along the tube by electromagnetic fields created in hundreds or thousands of accelerator cavities. The particles slam into a stationary target at the end of the tube. The limit to the beam energy a linear accelerator can reach is set by the cost of the cavities and their associated radio-frequency power supplies.

It was recognized early on that higher energies might be reached at lower cost by building a circular accelerator with only a few cavities, through which the particles would pass many times. The particles are kept on a circular trajectory by bending magnets interspersed with the accelerator cavities. Ultimately the beam is extracted from the ring and directed toward a target. The commonest circular accelerator today is the synchrotron.

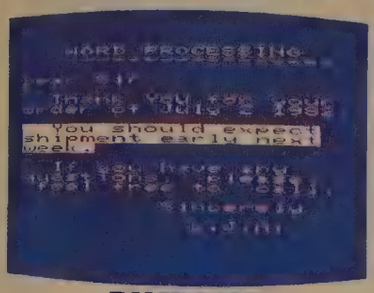
When a particle strikes a fixed target, much of the energy invested in accelerating the particle appears not as mass but as the kinetic energy of the product particles. If two particles collide head on, however, all their kinetic energy is available for conversion into mass. This is the principle of the storage ring.

In the simplest storage rings the colliding particles are electrons and their antiparticles, positrons. Because a particle and its antiparticle have the same mass but opposite charge, a single system of accelerator cavities and bending magnets can propel them around the ring in opposite directions. Where the counterrotating beams pass through each other, collisions are observed. The collision energy is the sum of the energies of the two beams. In the past 10 years or so some half-dozen electron-positron storage rings have been built. The largest of them has a collision energy of about 38 GeV and is being upgraded to 40 GeV, where particles incorporating a top quark might be accessible.

LEP, with electron and positron beams of 50 GeV each, will reach a collision energy of 100 GeV. Reaching that energy range is considered important because the mass of the Z^0 is predicted to be about 94 GeV. The Z^0 may be seen before LEP is completed in experiments now under way at CERN with a proton-antiproton storage ring. LEP, however, should give a much clearer view of the particle's properties. A second phase of the LEP project calls for the storage ring to be upgraded to a collision energy of 260 GeV. What will be seen at this energy is a matter of speculation, but it is entirely possible that new quarks and leptons will be discovered.

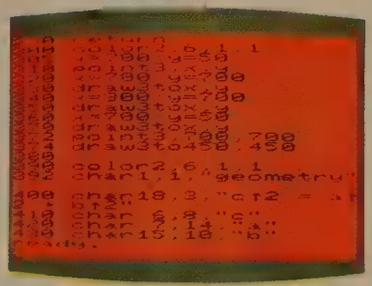
LEP will almost surely be the preeminent accelerator of the 1990's. Emilio Picasso of CERN, the director of the LEP program, has said that it will probably also be the last accelerator of its kind. The impediment to building a still more powerful synchrotron or storage ring is that charged particles following a curved trajectory emit the electromagnetic radiation called synchrotron radiation. The radiation dissipates energy that must be made up by supplying additional radio-frequency power. The energy loss increases as the fourth power of the beam energy. It can be reduced by building a ring with a larger circumference, but construction costs will rise.

IF YOU OWN A COMMODORE COMPUTER, YOU KNOW IT CAN DO ALL THIS.

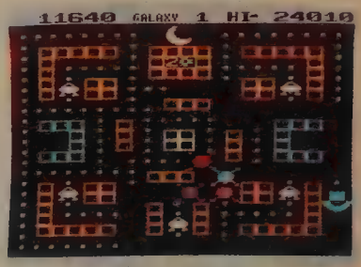


BUSINESS

EDUCATION

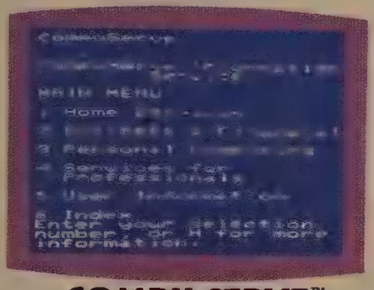


PROGRAMMING

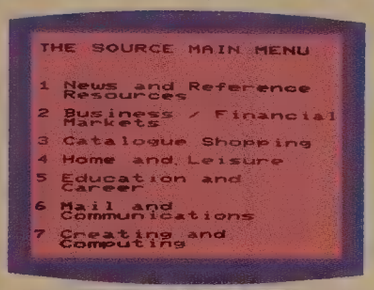


GAMES

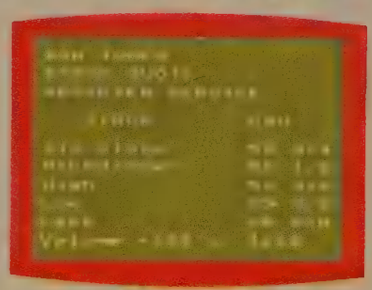
BUT DID YOU KNOW FOR ABOUT $100, YOU CAN ALSO GET IT TO DO ALL THIS?



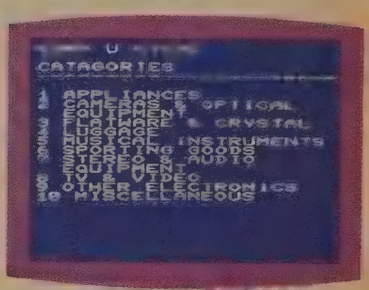
COMPU SERVE™



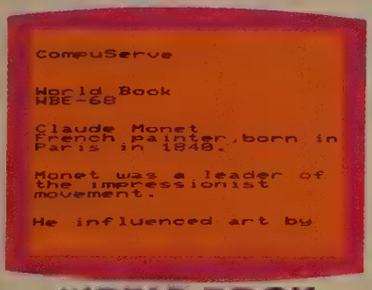
THE SOURCE™



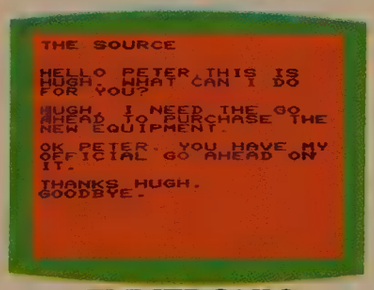
DOW JONES™



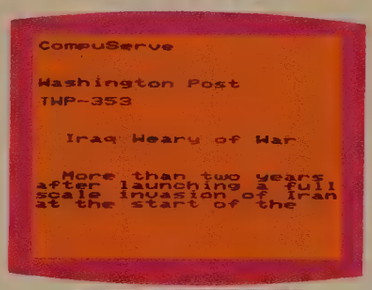
SHOP AT HOME



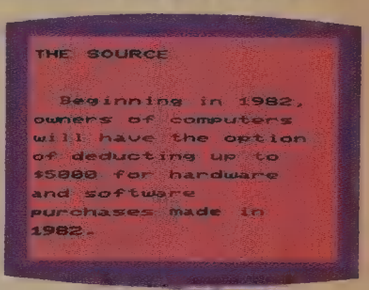
WORLD BOOK ENCYCLOPEDIA



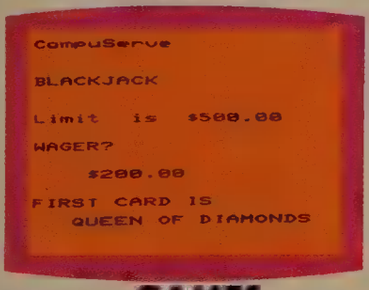
ELECTRONIC MAIL



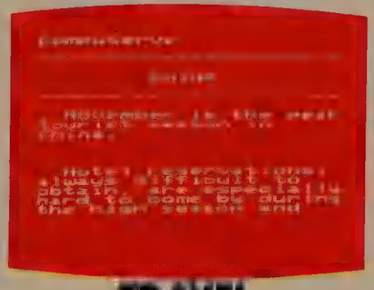
WIRE SERVICE NEWS



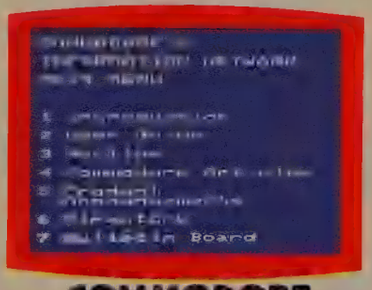
TAX ADVICE



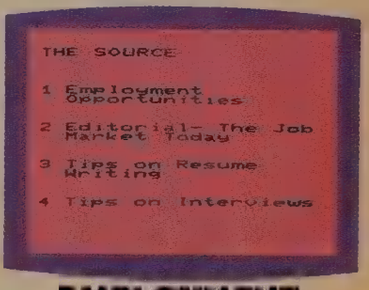
GAMES



TRAVEL INFORMATION



COMMODORE INFO. NETWORK



EMPLOYMENT OPPORTUNITIES

The screens at the top of the page show a few examples of how versatile the VIC 20™ or Commodore 64™ can be with the addition of Commodore software.

The screens below them give you a few examples of how much more versatile they can be with the addition of a Commodore VICMODEM.

For around $100, the Commodore VICMODEM will turn your VIC 20 or Commodore 64 computer into a telecomputer.

To make matters even better, Commodore includes a few little extras (such as a free hour's time on the two most popular telecomputing services) that add up to a value of $197.50* A nice return on an investment of about $100.

Most computer companies think it's reasonable to ask as much as $500 for a modem that'll give you telecomputing capabilities such as ours.

However, with a VICMODEM priced at around $100, we think we're being a lot more reasonable. Don't you agree?

commodore COMPUTER

U.S.A.–P.O. Box 500, Conshohocken, PA 19428; Canada– 3370 Pharmacy Avenue, Agincourt, Ontario, Canada M1W2K4 *Certain offers subject to change.
CompuServe is a trademark of CompuServe, Inc. and H.& R. Block Co. Dow Jones News/Retrieval Service is a registered trademark of Dow Jones & Co., Inc. The Source is a service mark of Source Telecomputing Corporation, a subsidiary of Reader's Digest Corporation, Inc.

There's a lot worth saving in this country.

Today more Americans who value the best of yesterday are working to extend the life of a special legacy.

Saving and using old buildings, warehouses, depots, ships, urban waterfront areas, and even neighborhoods makes good sense. Preservation saves valuable energy and materials. We can also appreciate the artistry of these quality structures.

The National Trust for Historic Preservation is helping to keep our architectural heritage alive for us and for our children.

Help preserve what's worth saving in your community. Contact the National Trust, P.O. Box 2800, Washington, D.C. 20013.

National Trust for Historic Preservation

Preservation builds the nation

Burton Richter of SLAC, who is directing the SLC project, argues that the cost of a circular collider is proportional to the square of the beam energy, whereas the cost of a linear collider is proportional to the energy itself. Therefore, he states, linear colliders will provide the cheapest way of reaching energies higher than the 260 GeV proposed for the second phase of LEP.

The SLC will be built by modifying the existing linear accelerator at SLAC, which is more than three kilometers long and already supplies electrons and positrons to two storage rings. A loop 600 meters in diameter is to be added at the end of the accelerator. Packets of electrons and positrons will be diverted into the loop, the electrons into one arc and the positrons into the other; they will collide on the far side. Electrons and positrons can be accelerated in the same direction by synchronizing their injection into the accelerator with the phase of the alternating radio-frequency fields in the cavities.

Since the SLC will provide a collision energy of about 100 GeV, it too could be used to examine the Z^0. It will be able to accommodate far fewer experiments than LEP, however, and there are no plans to upgrade it to higher energies. The main importance of the SLC may not be the physics experiments that could be done there. It is in itself an experiment in the design of a new type of particle accelerator, perhaps the first of a new generation.

Compatibility Genes

The histocompatibility antigens are molecules embedded in the cell's surface membrane that give an individual animal's tissue its own chemical identity. When a patch of skin or an organ is transplanted, it is these antigens that are recognized as nonself by the host's immune system, which proceeds to attack the foreign tissue. The rejection is most vigorous when the donor of tissue and the recipient have different versions of certain genes: those of the major histocompatibility complex, which in man occupies a short segment of chromosome 6 and is called the *HLA* region. In the mouse similar genes occupy the *H*-2 region of chromosome 17. The histocompatibility genes cannot have been conserved in evolution expressly to foil transplantation, and indeed they have been found to govern a wide range of biological functions. Among other things, it now appears, they determine characteristic body odors by which animals recognize one another.

In 1974 some staff members of the mouse-breeding facility at the Memorial Sloan-Kettering Cancer Center noticed that the social behavior of mice seemed to be influenced by their *H*-2 type. In a number of cages an inbred male mouse was kept with two females. One female was of the same inbred strain as the male and was therefore syngeneic, or genetically identical. The other female was *H*-2 congenic: genetically identical with the male except for the region of chromosome 17 carrying the *H*-2 genes. The male and the female whose *H*-2 genes were different tended to consort and nest with each other, excluding the syngeneic female. As it happened, Lewis Thomas, who was then president of Memorial Sloan-Kettering, had recently speculated that histocompatibility genes might give individual animals a characteristic scent. The combined stimulus of the breeding-trio observation and Thomas' suggestion led investigators at the center to join with a group at the Monell Chemical Senses Center in Philadelphia to learn whether mating preferences in mice are influenced by *H*-2 type, with recognition mediated by scent.

Their first experiment was a controlled test of the original breeding-trio observation. A male was caged with a syngeneic female and an *H*-2 congenic one. Thousands of trials confirmed that there is an overall bias in favor of mating between mice having different *H*-2 genes. Such pairing leads to *H*-2 heterozygosity in the offspring (that is, the offspring bear different sets of *H*-2 genes on their two copies of chromosome 17), presumably a beneficial trait.

The next step was to establish the mode of *H*-2 recognition, which seemed likely to be olfactory but could have been based on taste, or even on hearing since mice are known to communicate by ultrasonic signals. The ability of mice to distinguish between two *H*-2 types was tested in a *Y*-shaped maze. The test mouse was placed in the stem of the *Y*. Each arm of the *Y* was scented by air flowing through a chamber housing either a male mouse or its urine. The two odor-source mice were *H*-2 congenic with each other. Each thirsty test mouse was rewarded with a drop of water when it entered the arm of the maze scented by a mouse having a particular *H*-2 type (or by the urine of such a mouse). The experiment was strictly controlled to eliminate unknown adventitious signals and unintentional prompting by the operators of the maze. All the test animals, regardless of their genetic makeup, learned to distinguish between the *H*-2 types presented to them.

The *Y*-maze experiments have led to a number of conclusions, which were reported at a meeting of the International Society for Oncodevelopmental Biology and Medicine by Edward A. Boyse, Gary K. Beauchamp, Kunio Yamazaki, Judith Bard and Thomas. The sensory recognition of *H*-2 types is confirmed. Particularly in view of the successful training to distinguish urine alone, a prime mode of recognition (not neces-

To send a gift of Martell in the U.S., call 800-528-6148.
Imported by The Jos. Garneau Co., N.Y., N.Y. © 1981

INTRODUCING THE TRIM-SIZE

ALL NEW FORD

Trim new Ford 4-wheeler seats four in comfort. Split fold-down rear seats add extra versatility, plus up to 64 cu. ft. of cargo space.

Based on SAE specifications

V-6 engine standard, a V-6 with power the S-10 Blazer can't match. Extra pull for mud and snow. Big 23-gal. fuel tank.

Flip to 4-wheel drive from the driver's seat with auto-lock hub option. In snow or bad weather, shift to 4WD traction and relax.

BRONCO II

IT'S A BRAND NEW KICK!!

FORD

FORD DIVISION Ford

Built Ford tough, with husky frame and chassis. Solid Twin-Traction Beam front suspension. Proven manual or optional auto-lock hubs.

New sporty trim size makes driving and parking easy. Roomy comfortable interior. Plus independent front suspension for easy-rolling ride.

Perfect size for 4-wheeling—maneuverable Bronco II turns in less space than the S-10 Blazer or Jeep CJ's. Try it... for a brand-new kick!

Get it together—buckle up.

HAVE YOU HEARD?

The sound of hunger is louder than the rumble of an empty belly or the cry of a mother with nothing to feed her child. Hunger thunders through the generations and echoes against the dead end of abandoned dreams.

In the end, hunger can be heard in the scream of protest, in revolution and in rifle fire.

Oxfam America knows a better way. We work with people who are developing their own food and economic resources in 33 countries in Asia, Africa and Latin America. We put people to work in our own country, too, learning about the causes of hunger and what we, as responsible world citizens, can do about it.

We need your help to change the roar of hunger into a whisper of hope.

Oxfam America

Box N200
115 Broadway
Boston, Ma.02116
617 482-1211

A full, audited financial report is available from Oxfam America, or, for New York residents, from the Department of State, Office of Charities Registration, Albany, NY 12231.

sarily the only one) is olfaction. The recognition is not dependent on sex. Although mice choosing a mate seem generally to prefer animals of a different *H*-2 type, in the maze there is no obvious bias of untrained mice toward either a similar or a different type.

The workers at Monell and Memorial Sloan-Kettering are continuing their study, which opens many lines of inquiry. What other genes might define body odors? Already there is evidence that variation in the genetically allied region called *Qa:Tla* adjacent to the *H*-2 genes labels the urine of mice distinctively. What effects might the *H*-2 region have on aspects of reproductive behavior other than mating preference? There is evidence that it affects the phenomenon known as pregnancy block. The implantation of a fertilized ovum may be prevented when a female is exposed to a strange male soon after mating; there is now evidence that pregnancy block is more likely when the strange male is *H*-2-congenic, rather than syngeneic, with the stud male. Is there a relation between olfactory recognition and immunologic recognition? The major histocompatibility genes seem to be involved in both systems; olfaction, like the immune response, depends on cell-surface receptors, but the nature of the olfactory receptors and of the signals they transmit to the brain is still far from being understood.

Hybrid for Hybrid

A recent agricultural agreement shows that the transfer of technology between richer countries and poorer ones can go in both directions. A hybrid variety of rice developed in China and grown widely there is now under study in the U.S. In exchange a hybrid variety of cotton developed in the U.S. is being raised experimentally in China. The swap was arranged in 1980 by Armand Hammer of the Occidental Petroleum Corporation. The U.S. study is being done by Ring Around Products, Inc., an agricultural subsidiary of Occidental and the developer of the hybrid cotton.

The rice, a medium-grain glutinous variety known as Xian three-line, has given its American growers yields that are from 20 to 50 percent higher than those of commercial nonhybrid medium-grain varieties. In China, where six million hectares are now being planted exclusively with Xian three-line, it has reportedly increased overall rice production by some 13 million metric tons in the past five years. The plants are drought-resistant and suitable for a wide range of climates. According to the Chinese Science and Technology Commission, the grain has a protein content of from 9 to 11 percent.

The main impediment to commercial exploitation of the new rice is the cost of seed production, which is higher than it is for other hybrid crops such as corn. Growing seed for hybrid rice is a labor-intensive activity calling for considerable manipulation of the individual plants. Workers at Ring Around Products are looking for ways to mechanize seed production. A second objective is to further hybridize the variety so that it will bear nonglutinous long grains, the kind of rice preferred by American and European consumers, while preserving the Chinese hybrid's large root system and consequent resistance to drought.

Chilling Sound

Cycles of compression and rarefaction are the basis both of sound waves and of refrigeration machinery. Can a refrigerator be run by the sound from a loudspeaker? An acoustic heat engine devised by John C. Wheatley of the Los Alamos National Laboratory has demonstrated the principle.

If a piston is set in reciprocating motion in one end of a tube that is closed at the other end, heat builds up in the tube. Wheatley mounts a stack of thermally insulated fiberglass plates in such a tube and couples a loudspeaker to the piston. The loudspeaker emits sound at acoustic frequencies (from 150 to 1,000 hertz). Soon the end of the stack nearest the piston begins to cool while the other end heats up. Within a few minutes the difference in temperature can become quite large, perhaps 100 degrees Celsius.

Wheatley's experimental tube is a meter long and three centimeters in diameter. The plates are 10 centimeters long and are separated from one another by one millimeter. The tube is filled with a gas (in Wheatley's experiments it has been helium) and then closed with the piston. As the gas oscillates along the surface of each plate in response to the action of the piston, heat flows out of one end of the plate, through the gas along the surface of the plate and back into the other end. As it comes out of the cold end of the plate it is moving diffusively, that is, according to the laws of thermal conduction. As it moves along the plate it is flowing hydrodynamically, that is, because the gas is moving. It flows diffusively again as it passes back into the other end of the plate.

According to Wheatley, the acoustic heat engine can utilize a variety of thermally active materials. The thermal cycle depends on the configuration of the device and on the fact that thermal contact between the parts is imperfect. On making a practical machine, Wheatley says: "To get a large temperature difference you must reduce the heat input to the cold end. You can reduce external heat as much as you want, but a lot of heat is generated in the tube itself. We haven't tried yet to build something useful, but we're thinking about it."

All taste and no tin.

3M invented "Scotchtab" Closure Tape as a convenient opening for canned juices. That was over 20 billion cans ago.

Now we've improved on the idea. At the request of a Florida fruit and vegetable-juice packer, we created a tab that even protects the flavors of juices from the thin metal edge exposed when the hole is punched in the top. This helped our customer perfect a new kind of container to better serve its markets.

Listening to people has helped 3M pioneer over 900 products to solve industrial production and maintenance problems. We now make everything from tapes that hold tighter than nuts and bolts to floor coatings tough enough for duty on aircraft carrier flight decks.

And it all began by listening.

3M hears you...

For your free 3M Industrial Production Brochure, write: Department 090703/3M, P.O. Box 4039, St. Paul, MN 55104.

Name_____

Address_____

City_____ State____ Zip_____

Or call toll-free: **1-800-323-1718,** Operator 369. (Illinois residents call 1-800-942-8881.)

3M

WHERE DO BLACK DOCTORS COME FROM?

Nearly half of all black doctors practicing in the U.S. today are graduates of Meharry.

Sure, there are black medical students at Harvard and Stanford and Johns Hopkins. But no single private medical school in the U.S. produces as many black physicians as Meharry Medical College.

Ninety percent of our students require financial aid to some degree. To make more funds available for scholarships and student loans, and to maintain the quality of our training, our goal is to raise $27 million.

Meharry must depend largely on private and corporate contributions. We are eligible for only limited funding under Federal and State programs—and, as a professional school, receive no resources from the United Negro College Fund.

Won't you help? Because without Meharry, a lot of qualified youngsters couldn't even consider a medical career.

OPEN WIDE FOR MEHARRY
MEDICAL COLLEGE

Send your tax-deductible contribution, in any amount, to **Meharry Medical College, Second Century Fund, Nashville, TN 37208.**
The need is urgent, and your support is appreciated.

SUPERIOR STANDARDS IN MOLECULAR BIOLOGY AND BIOCHEMISTRY

Second Edition

Biochemistry

Lubert Stryer, *Stanford University*

Systematically developing each topic from the chemical principle to its biological and medicinal applications, the Second Edition of *Biochemistry* is organized around five major themes: conformation and dynamics; generation and storage of metabolic energy; biosynthesis of macromolecular precursors; information; and molecular physiology. The relationship between three-dimensional structure and biological activity is exemplified throughout in exceptionally clear, five-color illustrations. With many important additions — including treatment of 150 new topics, over 250 new illustrations, and a wealth of material demonstrating the power of biochemistry in elucidating physiological mechanisms — the Second Edition of *Biochemistry* is the definitive text for standard first courses in the field.

"[T]his is an excellent textbook for the general biochemist as well as for the medically oriented reader.... Though it is primarily a superb introduction, even experts will profit from its perspective and clarity of outlook." —*The New England Journal of Medicine*

1981, 949 pages, 983 illustrations
hardbound: ISBN 0-7167-1226-1 $32.95

DNA Replication

Arthur Kornberg, *Stanford University*

DNA Replication presents thorough coverage of the numerous DNA transactions that determine the structure and function of genetic material. Written by a Nobel laureate in medicine, this comprehensive study is augmented by illustrative tables and figures throughout, and provides a definitive account of knowledge and research through 1979.

"[T]his book is outstanding. Its concise, eloquent style makes it a delight to read, while the numerous excellent figures and tables and the comprehensive lists of references provide the reader with a wealth of relevant information." —*SEARCH*

1980, 724 pages, 308 illustrations
hardbound: ISBN 0-7167-1102-8 $38.95

1982 Supplement to DNA Replication

Arthur Kornberg, *Stanford University*

An essential addition to the study of DNA biochemistry, *1982 Supplement to DNA Replication* contains new figures, tables, and a review of important literature and developments in the field since 1979. The organization of topics in the *Supplement* parallels that of *DNA Replication*. Designed for use either independently or as a companion volume to Kornberg's earlier work, *1982 Supplement to DNA Replication* is indispensable reading for researchers and students in this rapidly developing field.

1982, 273 pages, 48 illustrations
paper: ISBN 0-7167-1410-8 $10.00

Second Edition

Physical Biochemistry
Applications to Biochemistry and Molecular Biology

David M. Freifelder, *Brandeis University*

A comprehensive introduction to important techniques of modern biochemistry and molecular biology, the Second Edition of *Physical Biochemistry* emphasizes the types of data yielded by these techniques and shows how these data can be interpreted. Theory is presented intuitively, and all problems have a strong experimental orientation. The text has been extensively revised and expanded to incorporate new developments in the field. Among the features of the Second Edition are an expanded section on macromolecules, presentation of the latest developments in autoradiography, inclusion of twice as many problems as in the First Edition, and treatment of numerous new subjects.

1982, 761 pages, 573 illustrations
hardbound: ISBN 0-7167-1315-2 $39.95
paper: ISBN 0-7167-1444-2 $21.95

Molecules to Living Cells
Readings from *Scientific American*

With Introductions by
Philip C. Hanawalt, *Stanford University*

Spanning the past four decades, this collection of twenty-six articles from *Scientific American* by distinguished scientists emphasizes molecular structures, their interactions in forming supramolecular structures, and the genetic control of their synthesis in the growing cells that constitute life. The book includes introductions providing background information, updating earlier articles, and placing scientific developments in perspective.

1980, 340 pages, 280 illustrations
hardbound: ISBN 0-7167-1208-3 $22.00
paper: ISBN 0-7167-1209-1 $10.95

To order copies of any of the above books, send a check for the total amount (please include $1.50 to cover postage and handling; California residents add appropriate sales tax) to: Promotion Department, Box MAG, W. H. Freeman and Company, 660 Market Street, San Francisco, CA 94104.

Mitochondrial DNA

The cell's energy-generating organelle has its own genetic system. The organization of its genes and even its version of the "universal" genetic code are different from those of the nucleus and bacteria

by Leslie A. Grivell

Some one and a half or two billion years ago, when the earth was still poor in oxygen, a primitive bacterium that made a precarious living from the anaerobic fermentation of organic molecules engulfed a smaller cell that had somehow evolved the ability to respire. The event was a turning point in organic evolution. Respiration liberates far more energy than fermentation, and the growing abundance of oxygen in the atmosphere must have been the driving force behind a symbiotic relation that developed between the two cells, with the aerobic cell generating energy in return for shelter and nutrients from its larger host.

In time the engulfed cell and others like it were to become subcellular organelles, passed on by host cells to their progeny. Eventually the host cells themselves changed, developing other subcellular structures and internal membranes and segregating their genetic material in chromosomes within a nucleus. These cells were the ancestors of all modern eukaryotic (nucleated) cells: protozoans and algae and the individual cells of fungi, plants and animals. The present-day descendants of those ancient symbiotic respiring bacteria are the mitochondria, the power plants of the eukaryotic cell.

Mitochondria are oval or worm-shaped organelles, about half a micrometer in diameter and from two to five micrometers long; this is roughly the size of many modern bacteria. The mitochondrion has an outer membrane and an extensively folded inner membrane that encloses a fluid matrix. The organelle is the site of oxidative phosphorylation, the primary source of cellular energy. In the fluid matrix organic molecules derived from the breakdown of foodstuffs are oxidized in a series of chemical reactions known as the citric acid cycle. Electrons removed in the course of oxidation are passed along a chain of respiratory-enzyme complexes arrayed in the inner membrane, driving the phosphorylation of adenosine diphosphate to form adenosine triphosphate (ATP), the universal energy carrier of cells. The cytoplasm of eukaryotic cells (the region outside the nucleus) contains from a few mitochondria to many hundreds; the greater the energy demands placed on the cell are, the more mitochondria it has.

Perhaps because they are descendants of a free-living bacterium, mitochondria have their own genetic material and the machinery to express it. The possibility that the organelles might have their own genes, distinct from the genome in the cell nucleus, was raised as early as 1949, when Boris Ephrussi found that the ability of baker's yeast to carry out oxidative phosphorylation seemed to be controlled by some factor in the cytoplasm rather than in the nucleus. It was only in 1966, however, that the first vertebrate mitochondrial DNA (that of the chick) was isolated and characterized. Since then the mitochondrial DNA's of many organisms have been under intensive study in a number of laboratories in Europe and the U.S. The intriguing questions have been: What does mitochondrial DNA do? How is it organized and expressed? Why should the cell bother to maintain a second genetic system alongside the one in the nucleus? Answers to the first two questions are emerging, and they make it possible at least to speculate about the broader third question.

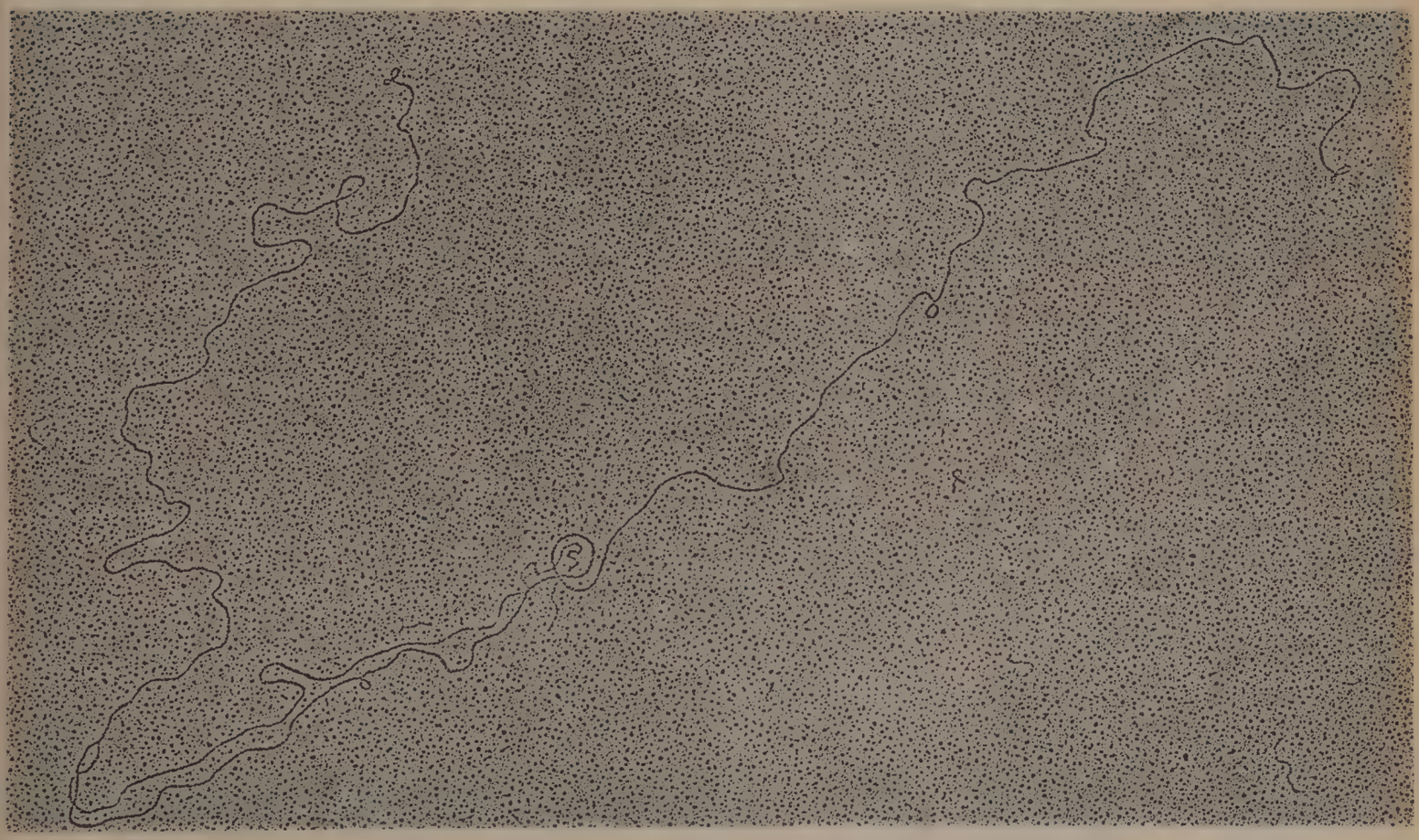
Important new data have been reported in the past two years, thanks largely to the application of molecular genetics and powerful new techniques for rapidly determining the sequence of nucleotides (the building blocks of DNA) in a genome. The complete nucleotide sequence of human mitochondrial DNA was published early in 1981 by B. G. Barrell and his colleagues in Frederick Sanger's laboratory at the Medical Research Council Laboratory of Molecular Biology in Cambridge; at the same time Giuseppe Attardi of the California Institute of Technology analyzed the transcription of human mitochondrial DNA into RNA. Soon the sequence of bovine mitochondrial DNA was published by Barrell's group and that of the mouse by David A. Clayton and his colleagues at Cal Tech. Meanwhile the sequence of a large part of the yeast mitochondrial genome has been unraveled by Alexander Tzagoloff of Columbia University and his group. What is becoming clear is that the diverse mitochondrial DNA's are unlike any other known DNA in several fundamental ways. Their unusual features may be the last traces of the primitive aerobic organism that long ago colonized fermenting bacteria and thereby initiated the eukaryotic revolution.

The Genome and Its Proteins

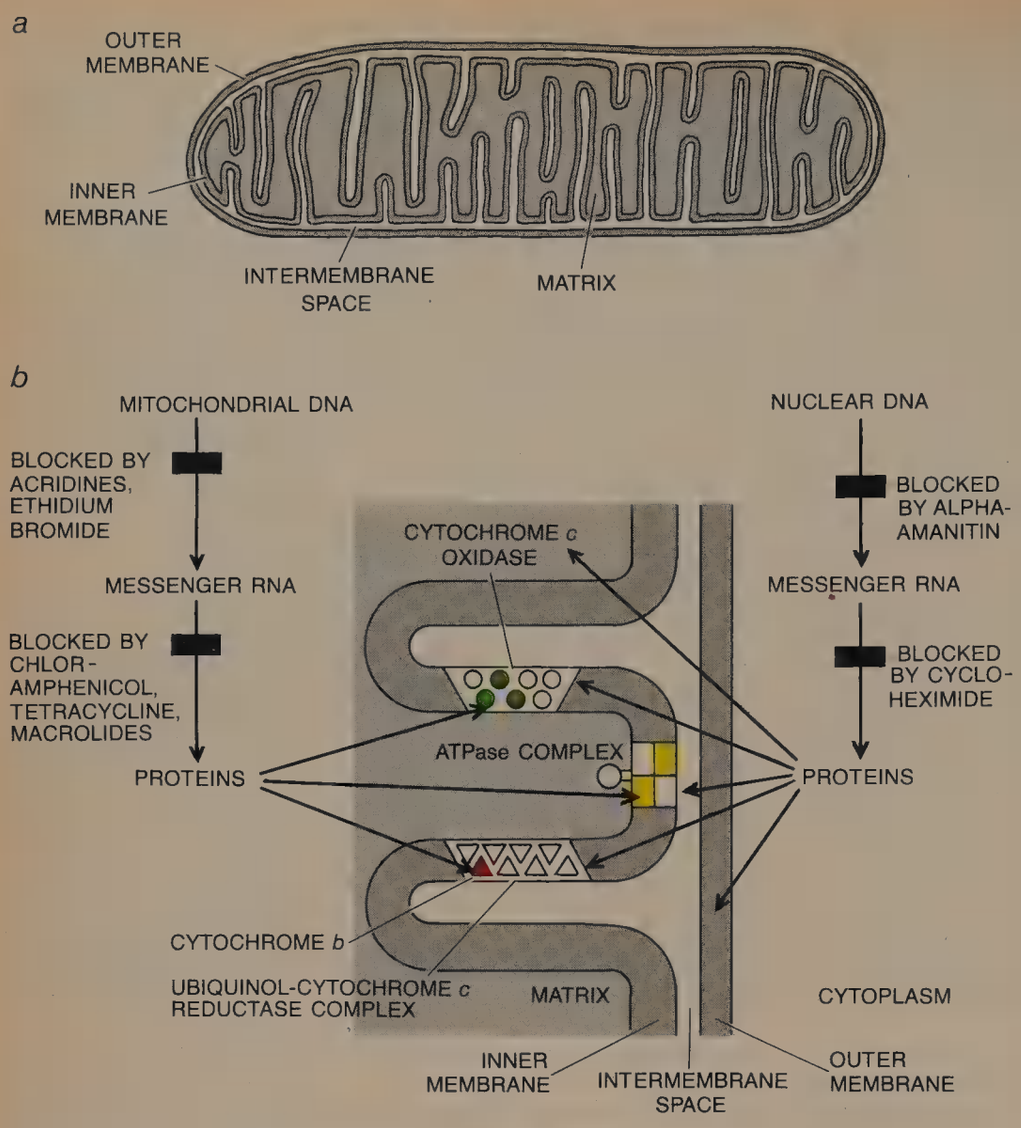
Mitochondrial DNA is like that of bacteria in that it consists of a single double helix of naked DNA, whereas the nuclear DNA of eukaryotes is tightly complexed with protein and is apportioned among a number of discrete chromosomes. Except in some protozoans the DNA helix is formed (again as in bacteria) into a circle. In most species, however, the mitochondrial genome is much smaller than that of the simplest bacteria; it is closer in size to the genome of many viruses. The size varies widely with the organism. Most animal mitochondria have only some 15,000 nucleotides, but in yeast the mitochondrial DNA is five times as long and in plants it is as much as five times longer still. Much of the difference in length between animal and yeast DNA's can be accounted for by the great difference in the way the same basic set of genes is organized. Whether this holds true for the mitochondrial DNA of plants is as yet not known.

Although the mitochondrion has its own genetic material, it is not genetically self-sufficient: most of the proteins that account for the structure and function of the organelle are encoded in the nucleus. The mitochondrion's own DNA specifies the RNA of its ribosomes (the jigs on which proteins are assembled) and its transfer RNA's (which carry specific amino acids to the

MITOCHONDRIAL DNA from the brewer's yeast *Saccharomyces carlsbergensis* is enlarged some 60,000 diameters in an electron micrograph made by Ernst F. J. Van Bruggen of the State University at Groningen in the Netherlands. The molecule is a "supercoiled" circle of double-strand DNA with a circumference of 26 micrometers. It is made up of some 75,000 nucleotides (the building blocks of DNA).



MITOCHONDRIAL GENOME of the protozoan *Tetrahymena pyriformis* is a linear molecule rather than a circle. The DNA is enlarged some 26,000 diameters in this micrograph made by Annika C. Arnberg and Van Bruggen. It is seen at an early stage in replication. A replication "eye" has formed near the middle of the molecule and is being extended as new DNA is synthesized along each of the strands.



STRUCTURE AND FUNCTION of the mitochondrion are defined jointly by genes in the cell nucleus and in the mitochondrion itself. The organelle, seen in longitudinal section (*a*), has ▪▪ outer membrane and ▪ deeply folded inner membrane that encloses ▪ fluid matrix. In all organisms examined so far the DNA of the mitochondrion encodes the organelle's own ribosomal RNA's and transfer RNA's, which are elements of the protein-synthesizing machinery. Other mitochondrial genes code for proteins. In yeast just six proteins are known to be specified by the mitochondrial genome (*b*): three subunits of the enzyme cytochrome *c* oxidase (*green*), two subunits of the enzyme ATPase (*yellow*) and the protein cytochrome *b* (*red*). All six take part in oxidative phosphorylation, the main source of cellular energy. Most proteins constituting the organelle or needed to operate its genetic system are encoded in the nucleus. It is possible to learn whether the nuclear genetic system or the mitochondrial one is responsible for ▪ particular synthesis by blocking one of the transcription or translation systems with a selective drug.

ribosomes for assembly into proteins). As for proteins, so far mitochondrial DNA's have been found to make messenger RNA that codes only for a few subunits of the respiratory-enzyme complexes in the inner membrane of the organelle. The rest of the inner-membrane proteins as well as those of the matrix (including the enzymes of the citric acid cycle) and those of the outer membrane are all encoded by nuclear genes; they are synthesized by ribosomes in the cytoplasm and imported into the mitochondrion. In addition to supplying the mitochondrion with most of its structural materials the nucleus encodes all the enzymes needed to replicate DNA and to transcribe it into RNA; the nucleus also encodes the proteins that combine with RNA to form the ribosomes and specifies other factors needed for the synthesis of mitochondrial proteins. In other words, the mitochondrial genetic system is maintained only through a considerable investment on the part of the nucleus and the cell's own protein-synthesizing machinery.

In the living cell the disparate contributions of the two genetic systems are coordinated, but there is some play in the coupling; the cell can live for a time while the genetic activity of either the nucleus or the mitochondria is blocked. Mitochondrial protein synthesis, for example, is blocked by antibacterial antibiotics such as chloramphenicol, tetracycline and the macrolide group; it is not affected by cycloheximide, which inhibits the activity of cytoplasmic ribosomes. One can arrange things so that proteins synthesized while one system is blocked are labeled with a radioactive isotope, and thus determine which system does what.

Labeling experiments have identified the products of mitochondrial DNA in a number of organisms. To identify the genes that encode the products and to map the positions of the genes on the circle of DNA was a harder task, however, or at least it was until the new methods for sequencing DNA were developed. The identification and mapping of genes by genetic analysis depends on noting and tracking mutations that interfere with a gene's activity without killing the organism. In the absence of oxygen a yeast can get by on fermentation alone, and so it can survive without functional mitochondria. It was therefore possible to note and characterize mutations in yeast that are lethal in other organisms, and as a result the mitochondrial DNA of yeast was the first to come under intensive study.

By now a combination of genetic analysis and DNA sequencing has provided inventories of the genes of both the human and the yeast mitochondrial genomes, together with detailed maps of their arrangement. Mitochondria from yeast and from man have essentially the same genes, but the organization of those genes is strikingly different.

Genome Organization

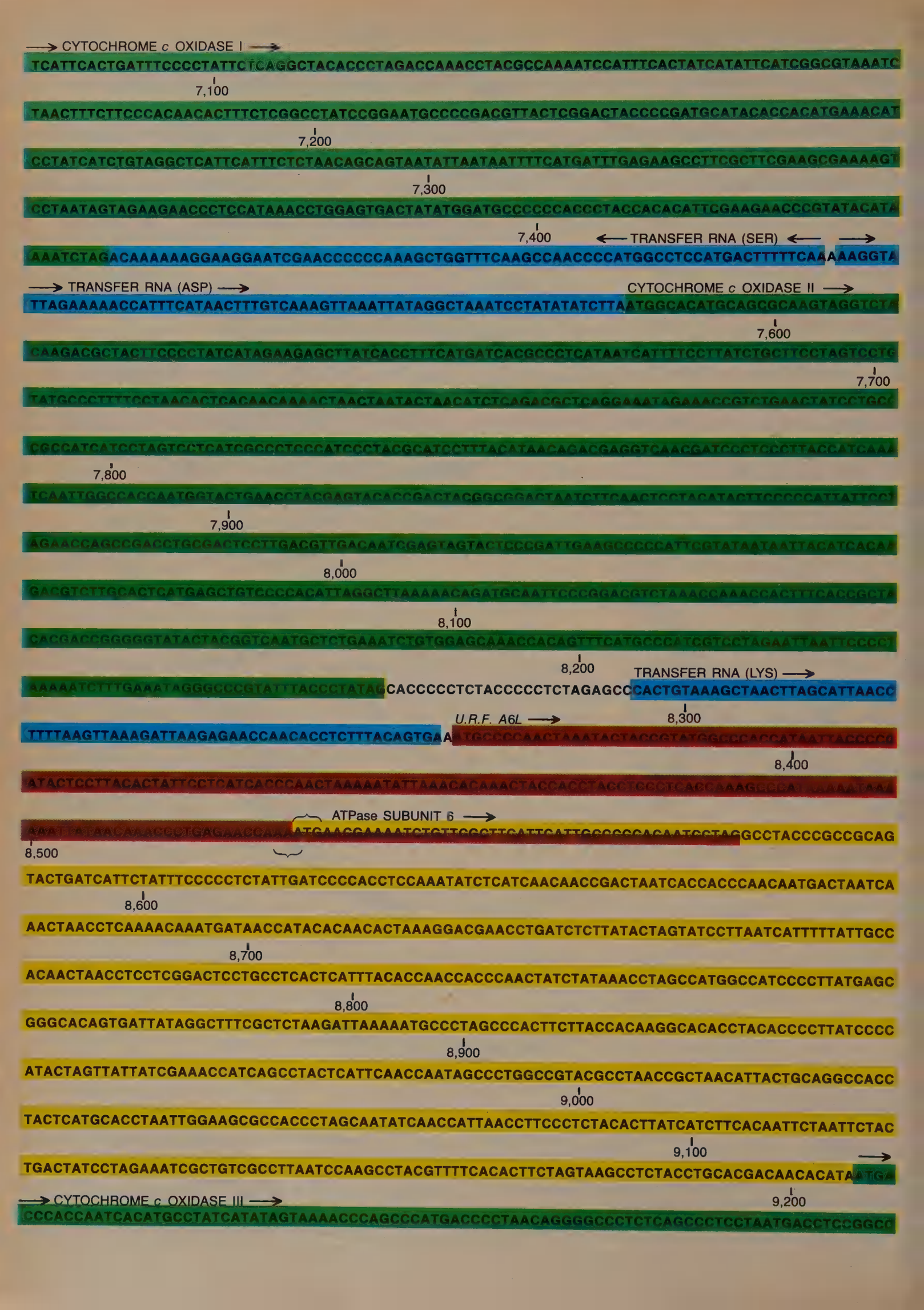
Each nucleotide of DNA is characterized by one of four bases, and the nucleotide sequences of the two strands of the double helix are complementary: adenine (*A*) on one strand pairs with thymine (*T*) on the other strand, and guanine (*G*) pairs with cytosine (*C*). Triplets of bases on one strand (the coding strand) serve as code words (codons) that, having been transcribed into RNA, specify either a particular amino acid or a signal for the initiation or termination of translation. In most eukaryotic chromosomal DNA the codons are a minority. The strings of triplets constituting a gene are flanked by long noncoding spacers and signaling regions that are not transcribed into RNA and by other stretches that are transcribed into messenger RNA but are not translated; the genes themselves are often split into pieces by noncoding intervening sequences ("introns").

Human mitochondrial DNA, in contrast, is a model of economy. It has only 16,569 base pairs, almost every one of which seems to have been put to use, with one strand or the other serving in turn as the coding strand. Indeed, the genes are so tightly packed that there are few if any noncoding nucleotides between them. There are genes for two ribosomal RNA's, for 22 transfer RNA's and for five membrane proteins; in addition there are eight unassigned reading frames. (These are regions where suc-

cessive triplets of nucleotides can be read as codons specifying amino acids and where the sequence of those triplets is not interrupted by "stop" codons. Such regions must specify a protein, but their protein product has not yet been identified.) None of the genes contain introns. Very little of the DNA codes for messenger RNA that is not subsequently translated into protein.

In both bacterial genomes and nonmitochondrial eukaryotic genomes the transcription of each gene is individually controlled by an interaction between transcribing enzymes and signaling sequences in noncoding regions of the DNA. The scarcity of such DNA in the human mitochondrial genome suggests transcription must be regulated in some unusual way. How this is managed is beginning to come clear as the result of painstaking studies by Attardi and his colleagues at Cal Tech. It appears there is only one major promoter, or initiation site for RNA synthesis, on each strand of human mitochondrial DNA. The primary product of transcription is therefore a full-length RNA copy of each strand. The two primary transcripts are then cleaved to yield the ribosomal RNA's, the transfer RNA's and the messenger RNA's that are subsequently translated into protein; the cleavage takes place even as the RNA is being synthesized. In the absence of noncoding DNA, what sites in the pri-

HUMAN AND YEAST MITOCHONDRIA have essentially the same set of genes, but their organization is different. The two DNA's are shown here as concentric double circles (one circle for each strand of the double helix) but at different scales: the yeast DNA molecule is actually about five times as long as the human one. The full nucleotide sequence of the human genome has been determined by B. G. Barrell and his colleagues; most of the sequence for the yeast *Saccharomyces cerevisiae* is also known, but some stretches remain unsequenced (*broken lines*). Colored segments show the extent of genes for known proteins (*green, yellow and red*), for ribosomal RNA's (*light blue*) and for transfer RNA's (*dark blue*), each of which is labeled with the abbreviation for its specific amino acid; the brown segments indicate unassigned reading frames (*U.R.F.*), which are presumably genes for proteins that have not been identified. The human genome is extremely compact, with little noncoding DNA between genes. In the yeast genome, on the other hand, there are long noncoding stretches. Several genes are interrupted by intervening sequences ("introns"), some of which in turn include unassigned reading frames.


CYTOCHROME c OXIDASE I
7,100
7,200
7,300
7,400
TRANSFER RNA (SER)
TRANSFER RNA (ASP)
CYTOCHROME c OXIDASE II
7,600
7,700
7,800
7,900
8,000
8,100
8,200
TRANSFER RNA (LYS)
U.R.F. A6L
8,300
8,400
ATPase SUBUNIT 6
8,500
8,600
8,700
8,800
8,900
9,000
9,100
CYTOCHROME c OXIDASE III
9,200

mary transcripts are recognized by the cleaving enzymes? Transfer-RNA transcripts flank just about every longer gene, and cleavage takes place precisely at the beginning and end of each transfer-RNA sequence. This suggests that the primary transcript folds in such a way that the distinctive shape of the transfer RNA's provides the recognition sites for the RNA-processing enzymes.

The lack of signaling space around the messenger RNA's similarly calls for unusual strategies to control translation. The signal for the initiation of translation is the same as the codon for the amino acid methionine. Nonmitochondrial messenger RNA's are preceded by sequences that serve as recognition sites for ribosomes, which bind there and thus label the following methionine codon as a "start" signal. Human mitochondrial ribosomes must be able somehow to recognize certain methionine codons as start codons without benefit of a preceding recognition site.

The problem of how to stop protein synthesis at the end of the gene is more complex. In nonmitochondrial messenger RNA's the termination of translation is signaled by one of several "stop" codons, among them *UAA* (*U* stands for uracil, the RNA analogue of thymine). Most of the human mitochondrial messengers either lack a stop codon (because they run into and overlap a following gene, usually one for a transfer RNA) or they have a stop codon that is truncated by the cleavage (mediated by transfer RNA's) described above. Those genes all end with either *U* or *UA*. How then is the stop codon supplied? The human mitochondrial messenger RNA, like most other eukaryotic messengers, is "polyadenylated," or given a long tail of adenine nucleotides, immediately after cleavage. The poly-*A* tail provides just what the translation machinery requires: a *UAA* stop signal.

Yeast mitochondrial DNA is as extravagant as its human counterpart is economical. The yeast genome has not yet been completely sequenced, but the DNA appears to code for two ribosomal RNA's, some 25 transfer RNA's, six membrane proteins and more than 10 unidentified proteins. In yeast (specifically in *Saccharomyces cerevisiae,* or baker's yeast) the genes are scattered over five times as much DNA. Coding sequences are separated by vast stretches of DNA consisting almost entirely of the bases *A* and *T*. At least three of the genes are split by introns. Long sequences at the beginning and end of genes are transcribed but not translated.

The Code Is Not Universal

Each codon in messenger RNA is "read" by the anticodon of a transfer RNA that brings a specific amino acid to take its place in a protein chain. The genetic code relating each codon of DNA (or its messenger-RNA transcript) to a particular amino acid is exactly the same in viruses, bacteria and animals from worms to man. It was thought to be universal, but it is not: mitochondria have their own version, which is subject to variation in different organisms. Mammalian mitochondria read *AGA* and *AGG* as stop signals rather than as codons for the amino acid arginine and translate *AUA* as methionine rather than as isoleucine; *AUA* and *AUU* can serve as start signals instead of *AUG*. In the mitochondria of the molds *Neurospora* and *Aspergillus,* on the other hand, *AGA, AGG* and *AUA* are translated according to the "universal" code. In yeast, as in mammals, *AUA* codes for methionine, but the four codons beginning with *CU* are read as threonine rather than as leucine. In the mitochondria of mammals, yeast, *Neurospora* and *Aspergillus* the amino acid tryptophan is specified by *UGA* as well as by the usual *UGG,* but in maize the extra codon for tryptophan appears to be *CGG*.

There are 61 codons (4^3, minus the three stop codons), but organisms make do with fewer transfer RNA's because a single transfer RNA can read more than one codon. It can do so because there is some freedom (or "wobble") in the pairing of the third base of a codon with certain bases in the first position of an anticodon. Under the wobble rules deduced some years ago by F. H. C. Crick, 32 transfer RNA's (but no fewer) should be able to read 61 codons. Yet sequence analysis shows that human mitochondrial DNA codes for only 22 transfer RNA's, yeast mitochondrial DNA for 25. No additional transfer RNA's seem to be imported from the cytoplasm, and yet 61 codons are read. How?

Each of the 22 transfer RNA's in mammalian mitochondria appears to be able to read a "family" of either two or four synonymous codons [*see illustration on page 86*]. The transfer RNA's for two-codon families seem to have conventional *G*: *U* wobble anticodons, that is, anticodons beginning with a *G* that is able to pair with *U* as well as with *C* or beginning with a *U* that can pair with *G* as well as with *A*. The transfer RNA's for four-codon families, however, ignore the Crick rules: they apparently read all four synonymous codons. Their anticodons have a *U* in the 5′ position. Either that *U* somehow pairs with all four of the 3′ codon bases or the anticodons read only the first two bases and ignore the last one. The possibility of such a "two out of three" reading process was suggested several years ago by Ulf Lagerkvist of the Faculty of Medicine of the University of Göteborg.

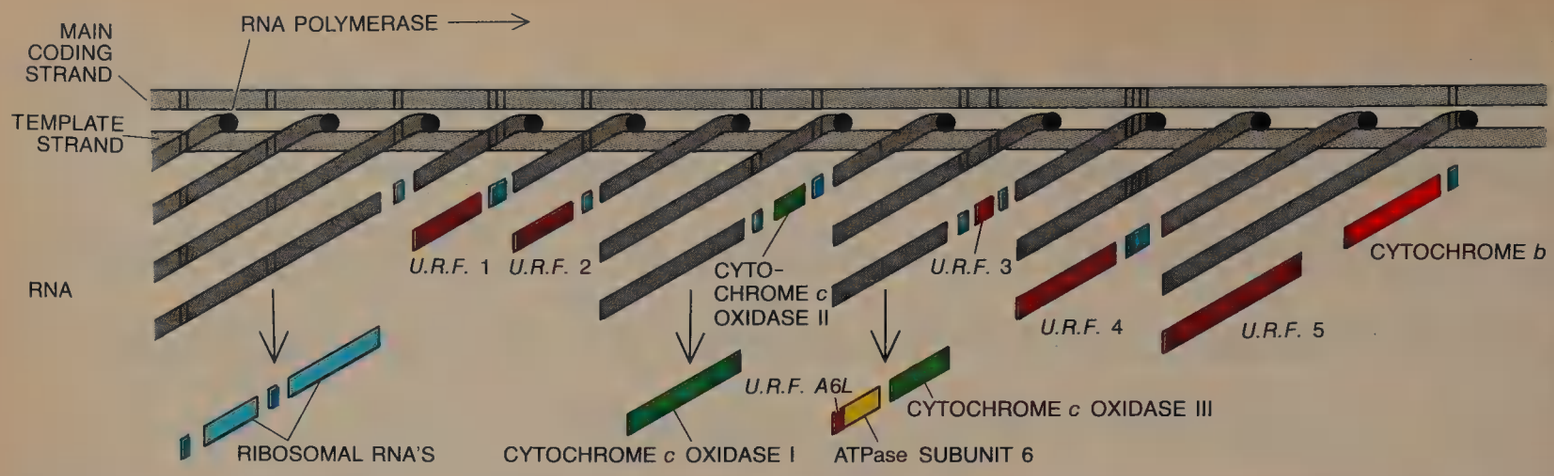
An intriguing question is how some transfer RNA's with *U* in the wobble position are held to reading two codons rather than four. What keeps them from recognizing codons ending in *U* or *C* as well as those ending in *G* or *A*? In the mitochondria of yeast and *Neurospora* and in many nonmitochondrial transfer RNA's such recognition is prevented by chemical modification of the wobble *U,* but so far sequence analysis of mitochondrial transfer RNA's in animals has not revealed any modification at this position. Most of these transfer RNA's do have an unconventional structure, however, and perhaps the structure influences the anticodon's specificity.

The Yeast Introns

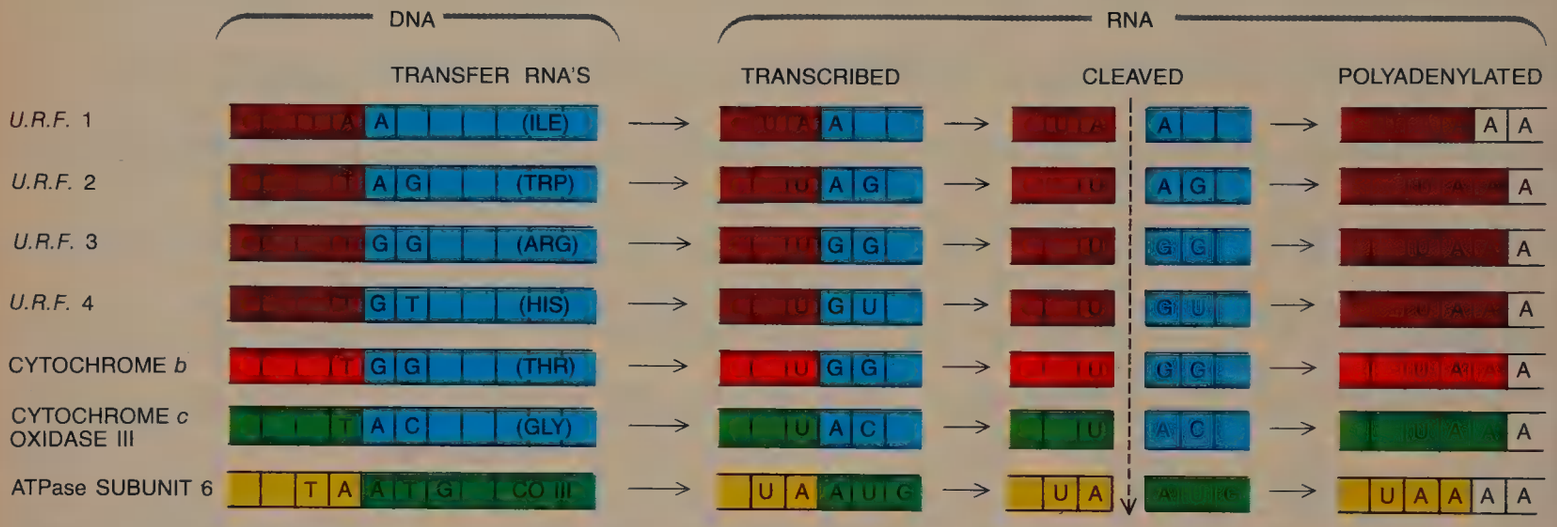
Split genes were first discovered in 1977, and by now it is clear that many of the chromosomal genes in mammals, birds and amphibians (and some of the genes in lower eukaryotic organisms) are split, with noncoding intervening sequences interrupting the "exons" coding for a protein. When such a gene is expressed, it is first transcribed, introns and all, into a precursor messenger RNA. This RNA is then "spliced," or processed: the introns are excised and the exons are joined to make a mature messenger RNA that is translated into protein. The introns of chromosomal genes, as far as anyone can tell, do not code for anything.

In yeast mitochondria things are different. Several years ago Piotr P. Slonimski and his collaborators at the Center for Molecular Genetics in Gif-sur-Yvette in France found that mutations in one intron of the yeast mitochondrial gene for cytochrome *b* (one of the respiratory enzymes) affect the splicing of the RNA transcribed from the gene. A series of elegant genetic experiments, coupled with DNA sequence analysis, revealed the reason: the splicing of the

SAMPLE OF THE NUCLEOTIDE SEQUENCE of human mitochondrial DNA is taken from the data gathered by Barrell and his colleagues. Only some 2,200 nucleotides of the 16,569-nucleotide molecule are shown here, but it is enough to suggest how tightly packed the genes are. In one region (beginning with nucleotide No. 8,527) two genes actually overlap, that is, the same nucleotides do double duty, encoding the end of U.R.F. *A6L* and the beginning of the gene for subunit 6 of ATPase. (The reading frame is different, as is indicated by the two brackets.) The sequence is the one for the strand that serves as the coding strand for most of the genes shown and that therefore has the same sequence as the corresponding messenger and transfer RNA's. The transfer RNA for the amino acid serine is actually specified by the complementary strand, which is transcribed in the opposite direction, as is indicated by its arrow.



HUMAN MITOCHONDRIAL DNA is transcribed into RNA and the RNA is simultaneously processed to form the mature ribosomal RNA's, transfer RNA's and messenger RNA's, according to Giuseppe Attardi of the California Institute of Technology. The diagram shows how the main coding strand is transcribed, beginning at a single major promoter site (*left*), into a single RNA transcript. The transcript is cleaved at the ends of the transfer-RNA sequences (*rectangles*) that flank almost every gene. The initial transcript and its successively processed segments are shown in gray. Mature RNA's are colored to match the scheme in the preceding three illustrations.



STOP CODONS signaling the termination of translation into protein are missing in seven human mitochondrial transcripts. In some the sequence runs into a downstream coding sequence; in others the three-nucleotide stop codon is cleaved, and thus truncated, when the RNA is processed. The necessary signal is supplied in an unconventional way by polyadenylation, a normal event in RNA processing in which a string of adenine nucleotides (*A*) is added to each transcript to make a messenger RNA. In this case the *A*'s fill out the stop codon *UAA*.

second intron of the gene is mediated by a protein that is specified in part by a segment of the intron itself!

Various forms of the cytochrome *b* gene have now been studied in different strains of yeast. The most complex form has six exons (designated *B*1 to *B*6) and five introns. Three of the introns include long unassigned reading frames. The entire gene is transcribed into a long primary transcript. Then intron No. 1 is excised. The effect of this step is to link exon *B*1 with exon *B*2 and at the same time to open up the reading frame in intron No. 2: to put its triplets in phase with those of the two preceding exons.

The spliced transcript of *B*1, *B*2 and the reading-frame portion of intron No. 2 now acts as a messenger RNA. It is translated into a protein whose amino-terminal head comprises the first 143 amino acids of cytochrome *b* (encoded by *B*1 and *B*2) and whose carboxy-terminal tail is encoded by the reading frame of the intron. This strange protein helps to catalyze the excision of intron No. 2. In so doing it destroys its own messenger RNA and thereby limits its own level in the mitochondrion. Slonimski calls this phenomenon "splicing homeostasis." He calls the protein encoded by the exon-intron combination a maturase to indicate that it may be not an enzyme itself able to catalyze splicing but rather a factor that somehow modifies the specificity of a preexisting splicing enzyme, probably one encoded by a nuclear gene.

Small Circular RNA's

The presence of an RNA maturase accounts neatly for the behavior of a number of mutant mitochondria having lesions within intron No. 2. They cannot process transcripts of the cytochrome *b* gene; instead they accumulate long RNA's from which only the first intron has been excised. They go on to generate large quantities of proteins that cross-react with antibodies to cytochrome *b* but are not cytochrome *b*. The synthesis of these proteins is blocked if mutations that terminate protein synthesis are introduced into exon *B*1. The proteins, then, must be defective maturases that arise from the translation of exons *B*1 and *B*2 and the long reading frame in intron No. 2; their overproduction is a direct result of their inability to destroy the messenger RNA that directed their synthesis.

Splicing homeostasis is probably not limited to the second intron of the cytochrome *b* gene. Introns No. 3 and No. 4 of that gene and the first four introns of the gene for subunit I of cytochrome *c* oxidase (another inner-membrane protein) also incorporate long open reading frames, and there is evidence that intron-specific maturases are involved in their processing.

Some mitochondrial splicing events generate unusual small circular RNA's, which were discovered by Annika C.

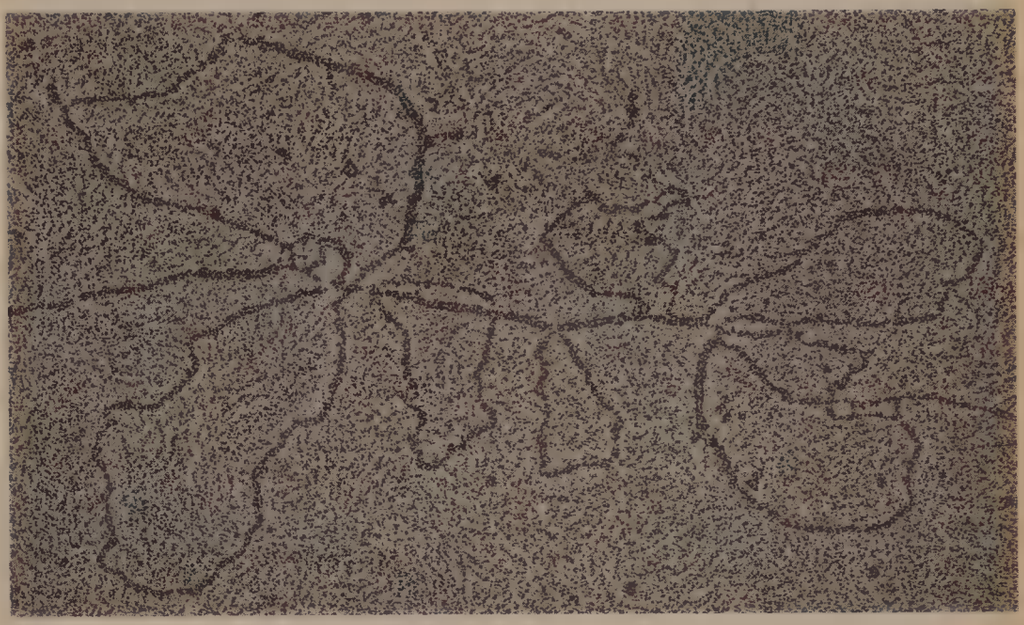
Arnberg of the State University at Groningen when she was collaborating with Gertjan van Ommen and me in electron-microscope studies to map mitochondrial transcripts. The circles turned out to be introns that had been spliced out of precursor RNA's and were then formed into closed loops, apparently by a covalent chemical bond. The circles have now been associated with three introns of the gene for subunit I of cytochrome *c* oxidase and with one intron (the No. 2 intron discussed above) of cytochrome *b*. The cytochrome *b* circle and one other from an intron with a long unassigned reading frame are apparently excised by mitochondrially encoded maturases, whereas the excision of the other two may depend on splicing enzymes encoded in the nucleus.

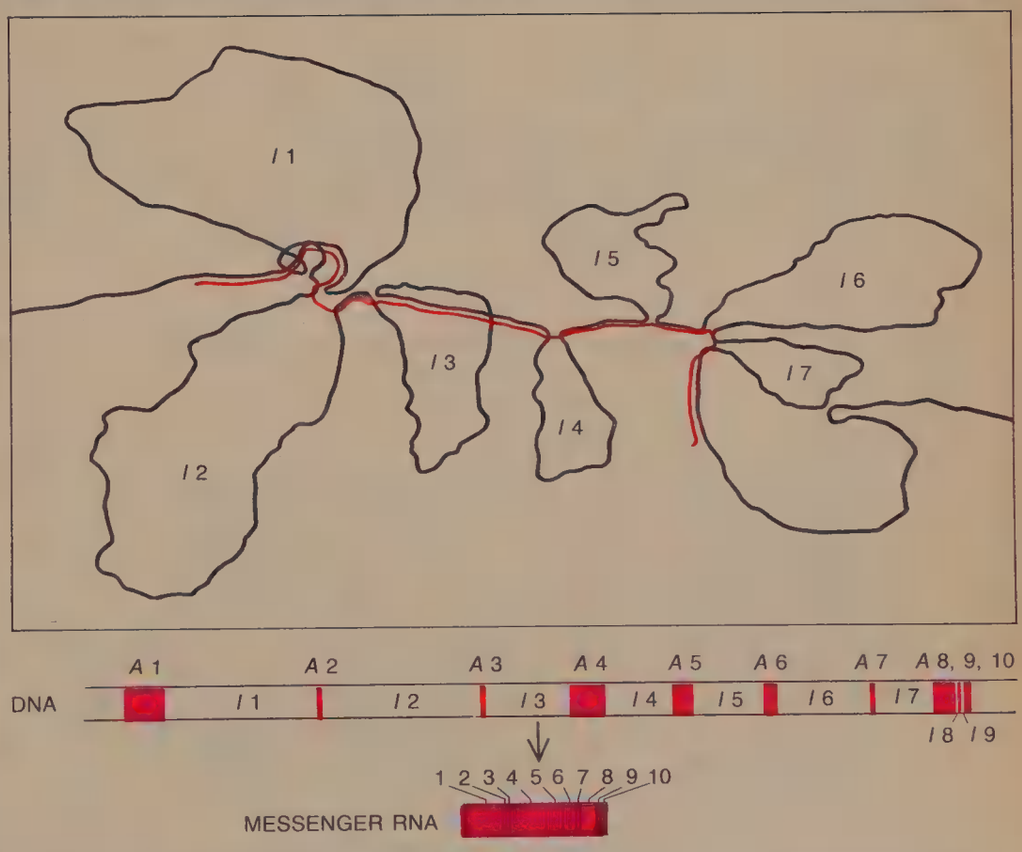
Crick once speculated that RNA circles might arise during splicing if a splicing enzyme can ligate the two ends of a single intron as well as join two exons. So far only one other instance has been reported where splicing gives rise to circular RNA's. In the nucleus of the protozoan *Tetrahymena,* Thomas R. Cech of the University of Colorado at Boulder and his colleagues reported a few years ago, an intron excised from the transcript of the gene for a ribosomal RNA is formed into a circle. When Cech's group tried to identify the enzyme that seemed to be mediating the excision and the circularization, they came to the surprising conclusion that there is no enzyme: the excision and circularization seem to depend on intrinsic properties of the RNA itself. The *Tetrahymena* RNA is the only example yet known of what Cech calls a ribozyme: a self-splicing RNA. Further study of mitochondrial circles that do not require maturases for excision should reveal whether ribozymes are implicated in their case as well.

Why Introns?

One remarkable property of the maturase associated with intron No. 2 is that it is dispensable. Yeast strains with shorter versions of the gene for cytochrome *b* are common; these genes often lack the first three introns, so that exons *B*1 through *B*4 are fused. The information originally present in the missing introns is not found elsewhere in the yeast mitochondrial genome, which suggests that the maturase is the mitochondrial equivalent of the worst kind of bureaucrat—one who exists only to regulate his own existence. Actually the three missing introns are members of a class of "optional" yeast mitochondrial introns defined by Johan Sanders when he was working in our laboratory at the University of Amsterdam a few years ago. There are other examples in the genes for the larger of the two ribosomal RNA's and for subunit I of cytochrome *c* oxidase. To classify an intron as optional one needs merely to find a strain that lacks it and yet is functional. Only a limited number of yeast strains have been examined so far; all yeast mitochondrial introns may eventually turn out to be optional. The optional nature of even a few introns puts some basic questions in a new light. Why are some genes split whereas others are not? Why should the same gene exist in both the split and the continuous form? Where do optional introns—and introns in general—come from?

Walter Gilbert of Harvard University





SPLIT GENE in yeast mitochondrial DNA is mapped by electron microscopy. The micrograph, made by Arnberg and Van Bruggen, and the interpretive map below it show ■ hybrid molecule: ■ single strand of DNA incorporating the gene for subunit I of cytochrome *c* oxidase (*black line in map*) has been allowed to hybridize with the messenger RNA (*color*) transcribed from it. In the DNA the sequences coding for the protein are interrupted by introns. Complementary sequences in the DNA and the RNA have paired to form double-strand regions; the introns, having no counterpart in the RNA, form single-strand loops. When such a gene is expressed, the primary RNA transcript is "spliced" to remove the introns, leaving only the "exons," or regions coding for the protein. Information from hybridization experiments has been combined with sequence data, gathered by Alexander Tzagoloff of Columbia University and by Lambert A. M. Hensgens in the author's laboratory, to reveal the organization of the exons (*A*1 through *A*10) and introns (*I*1 through *I*9) and of the mature messenger RNA after processing (*bottom*). (Introns *I*8 and *I*9 ■■ too small to be distinguished in a micrograph.)

has argued that in nuclear and viral genomes introns may confer a selective advantage by promoting more rapid evolution. By greatly increasing the length of a gene they increase the likelihood of various recombination events in which analogous segments of two alleles, or alternative versions of the same gene, are interchanged. Such recombination might bring two advantageous mutations, each in a different allele, together in one allele to create a doubly advantageous gene. It might also create completely new genes by what Gilbert calls exon shuffling: by bringing together, as exons of a single gene, several coding sequences that had previously specified different proteins. Moreover, a single-base mutation at a boundary between an intron and an exon can change the splicing pattern and thereby generate messenger RNA's encoding very different proteins. The changes in splicing induced by such mutations need not be absolute, so that the affected DNA becomes a "multiple choice" gene: it can code for both the original protein and the novel one depending on alternative splicing events, enabling the organism to test a solution without erasing the old one from its memory.

How did introns arise? The introduction of irrelevant sequences in an essential gene is ordinarily lethal. Even if an RNA-processing mechanism excises such sequences from the primary transcript of a gene, why should selective forces preserve a newly introduced intron? The cell derives no immediate advantage from the presence of an intervening sequence and certainly cannot foresee that it may be advantageous in the future. Such considerations led W. Ford Doolittle III of Dalhousie University in Nova Scotia to propose that introns were not introduced into eukaryotic cells but rather are part of their evolutionary heritage.

In Doolittle's view present-day prokaryotes (the bacteria) as well as the eukaryotic nucleus are likely to be derived from a primitive ancestral organism that was able to evolve rapidly because it had sloppy, unfaithful machinery for DNA replication and transcription and for translation. Such an organism's genetic information would have to have been encoded redundantly if any of it was to be expressed correctly. A genes-in-pieces organization might have been essential to ensure that transcripts of exons were at least occasionally assembled to direct the synthesis of functional proteins. As the genetic machinery became more accurate, Doolittle suggests, such insurance became less necessary; the replication and transcription of redundant and noninformational DNA became increasingly burdensome and there was pressure to eliminate it. The advantage of elimination was increased efficiency as exemplified by the streamlined genes of modern bacteria and the viruses that infect them. The price paid for this increased efficiency was the loss of a certain potential for further evolutionary change.

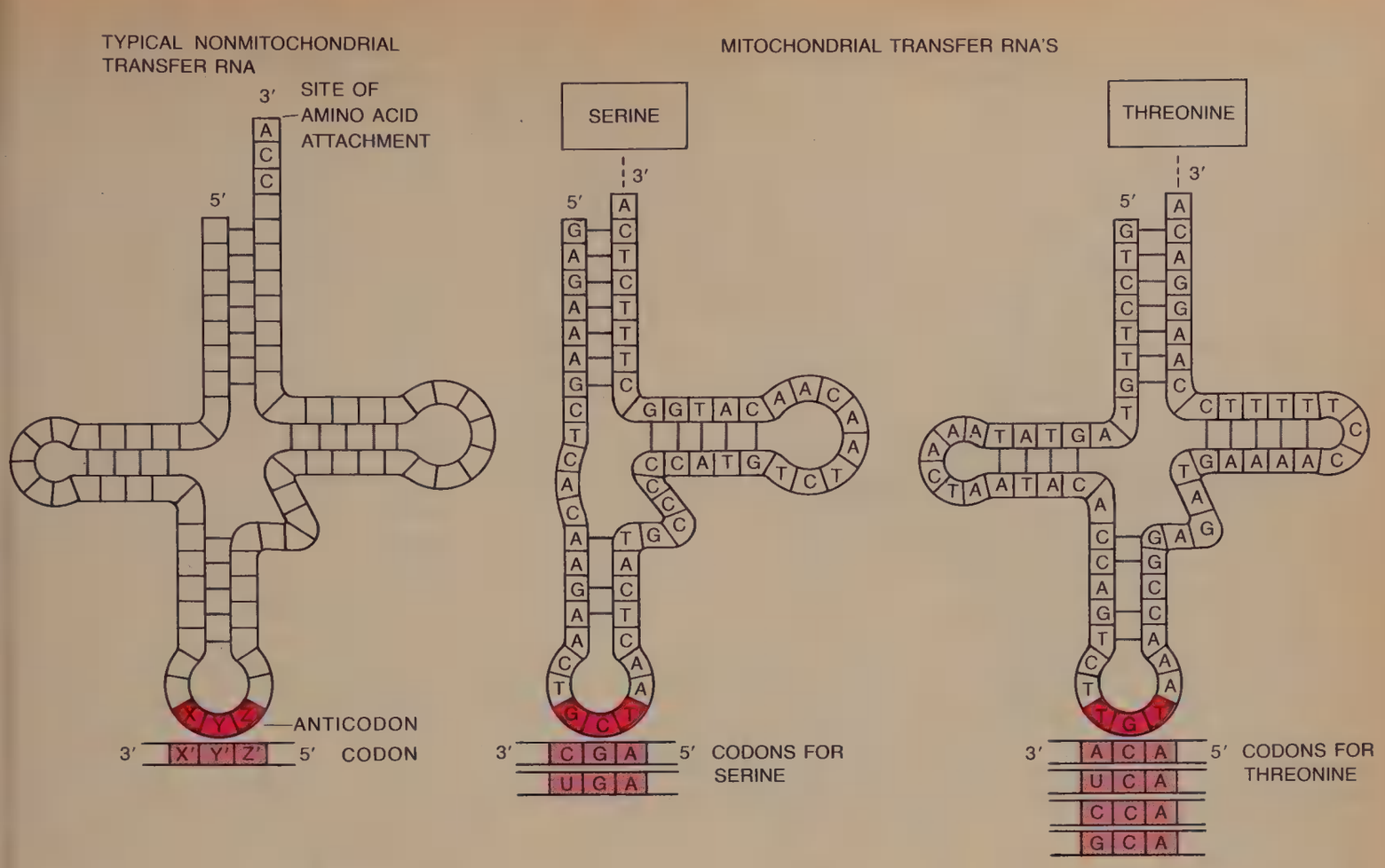
What about introns in mitochondrial genes? Many of the proposals advanced to explain introns in nuclear genes are unsatisfactory for mitochondrial ones, not only because they fail to explain the unassigned reading frames but also because their main predictions are not supported by the facts. Passive retention of introns with unassigned reading frames is not a good explanation because the cost to the cell of maintaining and expressing the many genes required to splice introns is appreciable, and such a complex system must be very vulnerable to disruption by mutation. Retention to facilitate recombination is an unlikely explanation because there is no indication that new mitochondrial genes have arisen recently by exon shuffling or that mitochondrially encoded proteins in fungi are evolving any faster than their counterparts in mammalian mitochondria, whose genes lack introns. Moreover, although exon shuffling is in principle an excellent way to spread mutations conferring advantageous characteristics (such as resistance to naturally occurring inhibitors of mitochondrial function), there is no reason to believe such mutations are fixed more efficiently in split genes than they are in continuous ones. Finally, there is no indication (although solid data are sparse) that any of the split genes in yeast mitochondria are multiple-choice genes.

FIRST LETTER	SECOND LETTER				THIRD LETTER
	U	C	A	G	
U	PHE	SER	TYR	CYS	U
	PHE	SER	TYR	CYS	C
	LEU	SER	*STOP*	(*STOP*) TRP	A
	LEU	SER	*STOP*	TRP	G
C	LEU	PRO	HIS	ARG	U
	LEU	PRO	HIS	ARG	C
	LEU	PRO	GLN	ARG	A
	LEU	PRO	GLN	ARG	G
A	ILE (MET)	THR	ASN	SER	U
	ILE	THR	ASN	SER	C
	(ILE) MET	THR	LYS	(ARG) *STOP*	A
	MET	THR	LYS	(ARG) *STOP*	G
G	VAL	ALA	ASP	GLY	U
	VAL	ALA	ASP	GLY	C
	VAL	ALA	GLU	GLY	A
	VAL	ALA	GLU	GLY	G

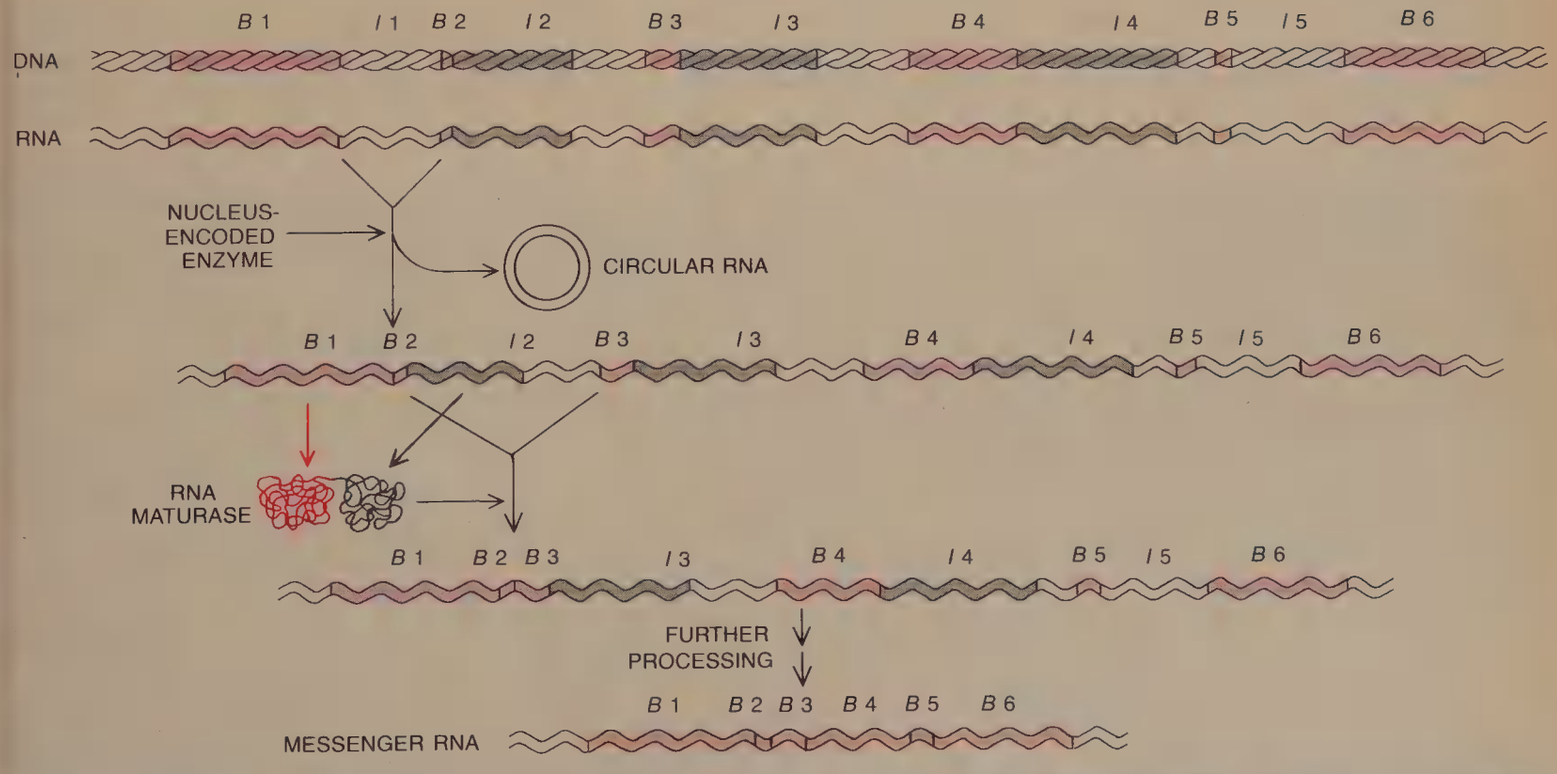
GENETIC CODE, in which a triplet of nucleotides (designated by their initials *U, C, A* and *G*) serves as a codon specifying a particular amino acid or a stop signal, was thought to be universal but is actually different in mitochondrial genetic systems. The table shows how codon assignments deviate in human mitochondria from those of the "universal" code; each mitochondrial deviation is shown in color, with the normal translation in brackets. Each of the 22 transfer RNA's in human mitochondria "reads" either two synonymous codons (*gray boxes*) or four (*colored boxes*). The codon response of the methionine transfer RNA's is not yet clear.

Introns and Repression

Perhaps, then, introns with unassigned reading frames have been retained in fungal mitochondrial DNA's because they confer some kind of selective advantage. What could the advantage be? One plausible explanation, put forward by Piet Borst of the University of Amsterdam, is that such introns may have a role in the fine regulation of mitochondrial-gene expression. Yeast cells can selectively repress the synthesis of certain mitochondrial proteins when the cells are growing by fermentation in the absence of oxygen, and some of the repression may be achieved at the level of RNA processing. Borst's proposal has the advantage of explaining why the repression is selective. Such genes as those for cytochrome *b* and subunit I of cytochrome *c* oxidase, which have several introns whose excision is likely to be controlled by a maturase, should be particularly sensitive to repression—and this has been found to be the case. The repression argument does not explain the presence of all yeast mitochondrial introns, however. One should not overlook the possibility that at least some of the introns without unassigned reading frames may not have any function at all. Given an effective mechanism for removing them from RNA, there may be little selective pressure to eliminate them from the DNA; indeed, mitochon-



TRANSFER RNA's generally have a stereotyped cloverleaf structure (*left*), the result of pairing (*gray lines*) between complementary nucleotide bases. (In RNA *A* pairs with *U*, and *G* pairs with *C*. In the mitochondrial transfer RNA's illustrated here the sequence is that of DNA rather than RNA, and so *A* is seen to pair with *T*.) Many mitochondrial transfer RNA's have an aberrant shape; the one for serine (*middle*) lacks a complete arm. A three-nucleotide anticodon on each transfer RNA reads a codon of messenger RNA. In some cases a transfer RNA can read more than one codon because there is some freedom, or "wobble," in the pairing between the first base of its anticodon and the third base of a codon; the *X* of the anticodon at the left might pair with a base other than *X′*. Many transfer RNA's in human mitochondria can apparently read as many as four codons (*right*), so that only 22 transfer RNA's suffice to read 61 codons.



RNA MATURASE, an enzyme specified in part by the second intron in the yeast mitochondrial gene for cytochrome *b*, helps to catalyze the excision of the intron encoding it. The gene has six exons (*B*1 through *B*6) and five introns. (Exons and introns are not drawn at the same scale.) It is transcribed into a precursor RNA, from which the first intron is promptly excised by a nucleus-encoded splicing enzyme. (Copies of the intron accumulate, surprisingly, in the form of single-strand circles.) The excision joins exons *B*1 and *B*2 to a long unassigned reading frame (*gray*) within intron *I*2. The RNA so assembled serves as a messenger RNA specifying the maturase, which is necessary for excision of *I*2; the maturase thus controls its own synthesis. There is evidence that other maturases help to excise *I*3 and *I*4.

dria may lack the machinery to do so efficiently. The introns could be a form of "selfish DNA," perpetuating itself for its own sake.

The fact that some introns in yeast, *Aspergillus* and *Neurospora* mitochondria are at closely corresponding sites in the genes they split suggests that the common ancestors of the genes may have had a similar organization. At least these introns may therefore directly reflect the genes-in-pieces organization of the primitive endosymbiotic organism from which mitochondria are thought to have evolved. Not all introns may be that old, however. Detailed comparisons of intron sequences in yeast, done by François Michel and Bernard Dujon of the Center for Molecular Genetics and by Linda Bonen of the University of Amsterdam, reveal many similarities both in the secondary structure of the RNA and in protein-coding sequences. These similarities strongly suggest that most present-day introns and their unassigned reading frames are members of a multigene family that originated in a single ancestral gene and has expanded by duplication and transposition in the course of evolution.

Such duplication with mobility is characteristic of the behavior of transposons: pieces of DNA that, in bacteria and in higher organisms, replicate and insert their copies at random sites in a genome. The closely similar yeast introns may possibly have arisen from transposons. If that is the case, it could explain the origin of the maturases. Most bacterial transposons encode a transposase: an enzyme that catalyzes the insertion of the transposon. Transposition involves cleavage precisely at the two ends of the transposon, just as RNA splicing requires precise cuts at the ends of an intron. The maturases could, therefore, have evolved from transposases.

Why Mitochondrial DNA?

Why should a separate genetic system be necessary for just one cellular organelle among many? The mitochondrial system cannot be explained as a form of gene amplification, necessary to provide the mitochondrion with large quantities of certain proteins. High levels of protein synthesis can be achieved by other means, and if amplification were necessary, it could as well take place in the nucleus. Nor can mitochondrial DNA be explained as the necessary source of proteins that are too insoluble to be transported to their site in the inner membrane. Subunit 9 of the ATPase complex, one of the most hydrophobic of the proteins specified in yeast by mitochondrial DNA, is specified by a nuclear gene in *Neurospora* and in mammals, synthesized on ribosomes in the cytoplasm and imported into the mitochondria.

If a separate mitochondrial genetic system is not necessary, why is it there? Borst thinks mitochondrial DNA is simply an evolutionary dead end, a relic of the putative symbiotic bacterium. According to Borst's hypothesis, in the early stages of evolution there were few barriers to the transfer of information between the genetic system of the endosymbiont and that of the host, and gradually most of the endosymbiont's genes became part of the host cell's genome. As the eukaryotic cell became more sophisticated (as its DNA was organized in chromosomes and segregated in a nucleus) and perhaps as the nuclear and the mitochondrial genetic codes diverged, the likelihood of transfer decreased; finally it became impossible. A few genes remained isolated in the mitochondrion. Being essential, they persist.

The hypothesis is attractively simple. It accounts nicely for the fact that a number of mitochondrial enzymes encoded in the nucleus have similarities to corresponding enzymes in bacteria. It predicts that some genes will be found in the mitochondrial DNA of one organism but in the nuclear DNA of another, as in the case of the gene for subunit 9 of ATPase. It allows for intermediate situations in which both mitochondrial and nuclear versions of the same gene are present in the same cell. That may be true in *Neurospora,* in which Paul van den Boogaart and Étienne Agsteribbe of the State University at Groningen have found copies of the gene for subunit 9 of ATPase in both mitochondrial and nuclear DNA.

Borst's hypothesis predicts further that some mitochondrial DNA's may have more genes than are in the basic set present in both yeast and mammalian mitochondria. The mitochondria of higher plants have extremely large genomes, ranging from perhaps 250,000 nucleotides in maize to as many as 2.4 million in the squash and melon family. Much of the difference may be due to the presence of noncoding DNA, but laboratory experiments show that plant mitochondria synthesize more proteins than their fungal or mammalian counterparts do, and there is good evidence that in plants the phenomenon of male

UNUSUAL MITOCHONDRIAL DNA of the protozoan *Trypanosoma brucei* is enlarged some 16,000 diameters in an electron micrograph made by J. H. Hoeimakers of the University of Amsterdam. The DNA is in the act of segregation: having been replicated, it is dividing to form two genomes when the organelle divides. The DNA is a network of interlocked maxicircles (*black arrow*) and minicircles (*white arrow*). The maxicircles seem to be equivalent to mitochondrial DNA in other organisms; function of minicircles is not known.

sterility (inability to produce fertile pollen) can be the result of changes in mitochondrial DNA.

Evolution toward Diversity

Whereas the major products of mitochondrial genes appear to have been largely conserved in the course of evolution, the organization and the mode of expression of those genes have not been conserved. The diversity of organization is nowhere better exemplified than in the mitochondrial DNA of *Trypanosoma,* a genus of unicellular protozoans that includes parasites of man and domestic animals. The DNA is an enormous network of linked circles—perhaps 10,000 of them. There are maxicircles ranging in length from some 20,000 to 36,000 nucleotides. They have regions of homology with pieces of yeast DNA and are transcribed; they are likely to be the counterpart of the mitochondrial DNA in other organisms. Most of the DNA, however, consists of minicircles less than 3,000 nucleotides long. These are not transcribed, and they have undergone evolution at a high rate, which is characteristic of noncoding DNA. The function of the vast network of minicircles is not known, nor is it at all clear why trypanosomal mitochondrial DNA should have such an unusual structure.

This final visit to one of the outposts of biology will, I hope, convince the reader that mitochondrial genetic systems are diverse. Their variety is the result of a high rate of nucleotide-sequence evolution—perhaps upward of 10 times the rate for genes in the nucleus. The apparent lack of restraint on sequence evolution may be related to the simplicity of the system. Its function is primarily to synthesize a limited number of proteins in rather constant amounts, and it may be able to accommodate mutations that would be deleterious in a system geared to the differential regulation of many genes. It is possible too that in each organism different selective pressures have pushed the mitochondrial DNA's into different and sometimes extreme solutions to the problem of gene expression.

Finding out what these pressures may be is one task for the future. Another is to learn through study of this limited system just what the minimal requirements of a functional genetic system are. Better understanding of the mitochondrial genome should reveal some of the evolutionary pathways leading to mechanisms of gene expression in present-day bacteria and eukaryotes. To judge from what the study of mitochondria has already revealed, the lessons to be learned about the organization and the expression of mitochondrial DNA should have applications extending far beyond the boundary membrane of this modest organelle.

PARALLEL PROCESSING

... AND YOUR CAREER.

Compiler Designers, SPERRY Computer Systems is developing a new line of scientific processors to be the base for the most cost-effective scientific computing system in the industry. The FORTRAN COMPILER GROUP is developing a new set of scientific compilers for these new processors.

If you have experience in compiler design for any parallel/ pipeline processors and/or a Masters or Doctorate level degree in Computer Science with emphasis in state-of-the-art compiler techniques for scientific processing, we would be interested in discussing your career possibilities at SPERRY.

Interested individuals can call or send resume to Char Nelson at SPERRY Computer Systems, P.O. Box 43942, MS 4973 SA, St. Paul, Minnesota 55113. (612) 631-5785. An equal opportunity employer M/F/V/H.

SPERRY

COMPUTER SYSTEMS

We understand how important it is to listen.

The Future of the Universe

A forecast for the expanding universe through the year 10^{100}: All protons will decay, galaxies will form black holes and black holes will "evaporate." If the universe collapses, it may cycle

by Duane A. Dicus, John R. Letaw, Doris C. Teplitz and Vigdor L. Teplitz

In the past several years developments in the physics of high-energy interactions of elementary particles have stimulated a remarkable growth in cosmology. The attempt to describe all the fundamental forces of nature as different aspects of one underlying force has had partial success in the "grand unified theories" of elementary particles. Through such theories it is possible to give tentative descriptions of dominant physical processes from extremely low temperatures to temperatures on the order of 10^{32} degrees Kelvin. The properties of matter at densities of cosmological interest, which vary from less than 10^{-300} gram to more than 10^{100} grams per cubic centimeter, can thereby be addressed. The extreme conditions at the ends of these scales can prevail only in the very early or the very late stages in the evolution of the universe.

A few years ago cosmologists began to show how theoretical physics can give a consistent account of the history of the universe to within 10^{-35} second of its beginning. More recently several physicists and cosmologists, including the four of us, have extrapolated cosmic events into a future that is up to 10^{100} times the current age of the universe. Apart from the intrinsic fascination of such an enterprise, there are several reasons for undertaking it. From the standpoint of theoretical physics the extrapolation makes possible a kind of thought experiment: By determining how the diverse effects predicted by a theory interact at a given time, one can test the overall consistency and plausibility of the theory. No ordinary laboratory is ever likely to attain the temperatures and densities at which grand unified theories make predictions, and so the theories are being verified in what some physicists call the ultimate laboratory, namely the universe as a whole. From the standpoint of cosmology the importance of our extrapolations is that the grand unified theories also have consequences that can be tested in terrestrial laboratories, and so the predictions they make under extreme conditions can be confirmed. One can therefore delineate a sequence of events in the distant future with considerably more detail than has previously been possible.

The framework for our calculations about the remote future is the big-bang model. There is now general agreement that the evolution of the universe must have been determined in the first few moments of the big bang. According to the model, the universe began as the explosion of a highly compact, primordial entity some 10 to 20 billion years ago. The term big bang and the terminology associated with an explosion are appropriate in part, because matter and energy do indeed appear to have been flung outward in space from a common origin. The term can also be misleading, however, because it is not correct to suppose the expansion can be visualized from the outside. There is no external vantage because what is exploding is the entire universe. Space itself is expanding in the sense that the separation between any two galaxies is growing at a rate that depends on their separation. Viewed from our galaxy all other galaxies appear to be receding at a rate that increases by 17 kilometers per second for every million light-years' separation. Mathematically the big bang is described by Einstein's general theory of relativity.

The recession of distant galaxies is inferred from the observed shift of their radiation toward the red end of the electromagnetic spectrum. The expansion and cooling of the universe has altered even more dramatically the spectrum of another kind of radiation, namely the cosmic microwave background radiation, discovered in 1964 by Arno A. Penzias and Robert W. Wilson of Bell Laboratories. The background radiation seems to be coming with equal intensity from every point in the sky; the intensity varies with direction by less than one part in 10,000. It was the scattering and annihilation of electrically charged particles in the early stages of the big bang that gave rise to the background radiation. The radiation was therefore once very hot, but it has since been cooled by the expansion of the universe to a temperature of about three degrees K.

Nuclear reactions at the density and temperature of the universe about three minutes after the start of the big bang were responsible for the synthesis of helium and, to a much lesser extent, other light elements. The universe cooled too quickly, however, for carbon and heavier elements to form, and a generous supply of hydrogen was left over to serve as nuclear fuel for stars. Earlier, reactions only 10^{-38} second after the universe began may have generated the observed excess of matter over antimatter, in which case most grand unified theories predict the eventual decay of all nuclear matter. The current understanding of the conditions in the early universe has not, however, been sufficient to answer the most compelling question in cosmology: Will the universe continue to expand forever, or will gravitational forces halt the expansion and drag all space and time back into the fireball with which the universe began?

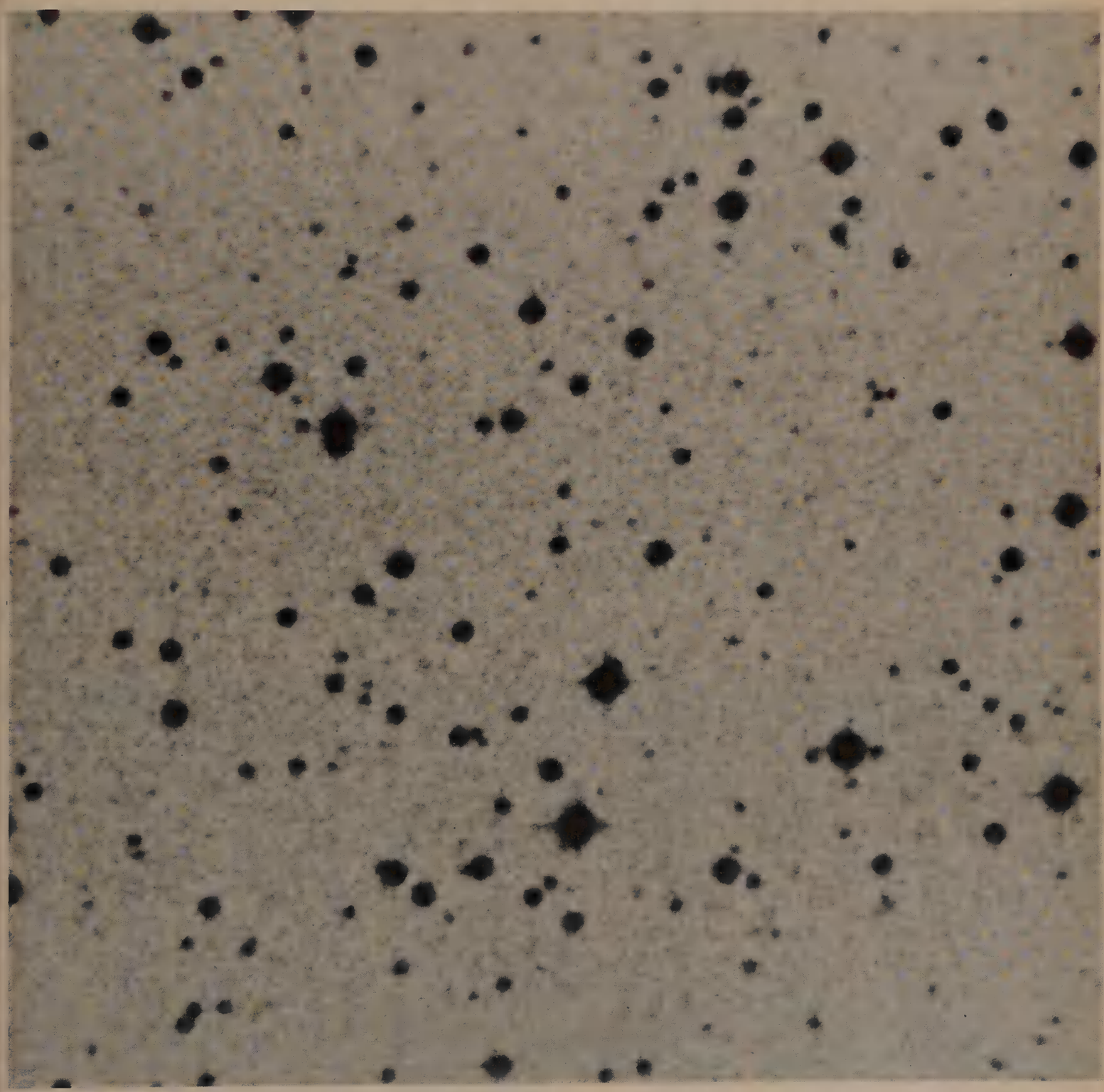
Although many years of observational effort and theoretical consideration have been given to the last question, the issue is still undecided. In principle the simplest way to resolve it is to measure the amount of matter in a large, representative volume of space and so estimate the density of matter in the universe at large. If the density is equal to or greater than a critical density, which is about three protons for every 1,000 liters of space, gravity will eventually overcome the expansion and the universe will collapse. A universe of this kind is said to be closed. If the density is less than the critical value, the universe will remain open; the expansion will be slowed by gravity but never stopped. Note that physical processes such as stellar burning or the decay of elemen-

tary particles cannot change a closed universe into an open one.

Estimates of the number of stars in a typical galaxy and the number of galaxies per unit volume give a value for the density of matter that is only 5 percent of the critical density. On this basis one might conclude that the universe is open and will continue its expansion forever. An estimate of mass based on the counting of stars, however, includes only luminous matter; if nonluminous matter accounts for a significant fraction of the mass of the universe, the estimate is not accurate.

There is some evidence of nonluminous matter in galaxies. The speed with which a star revolves about the center of its galaxy should increase with distance from the galactic center for stars lying in the interior of the galaxy and should decrease with distance if the star is at the periphery. For outlying stars, however, the expected decrease in speed of revolution is not observed; the observed distribution of speeds leads to the conclusion that at the periphery of a galaxy there is a halo of nonluminous matter whose total mass is one or two times that of the luminous galactic matter. Hence nonluminous matter within galaxies may account for from 5 to 10 percent of the mass needed to close the universe.

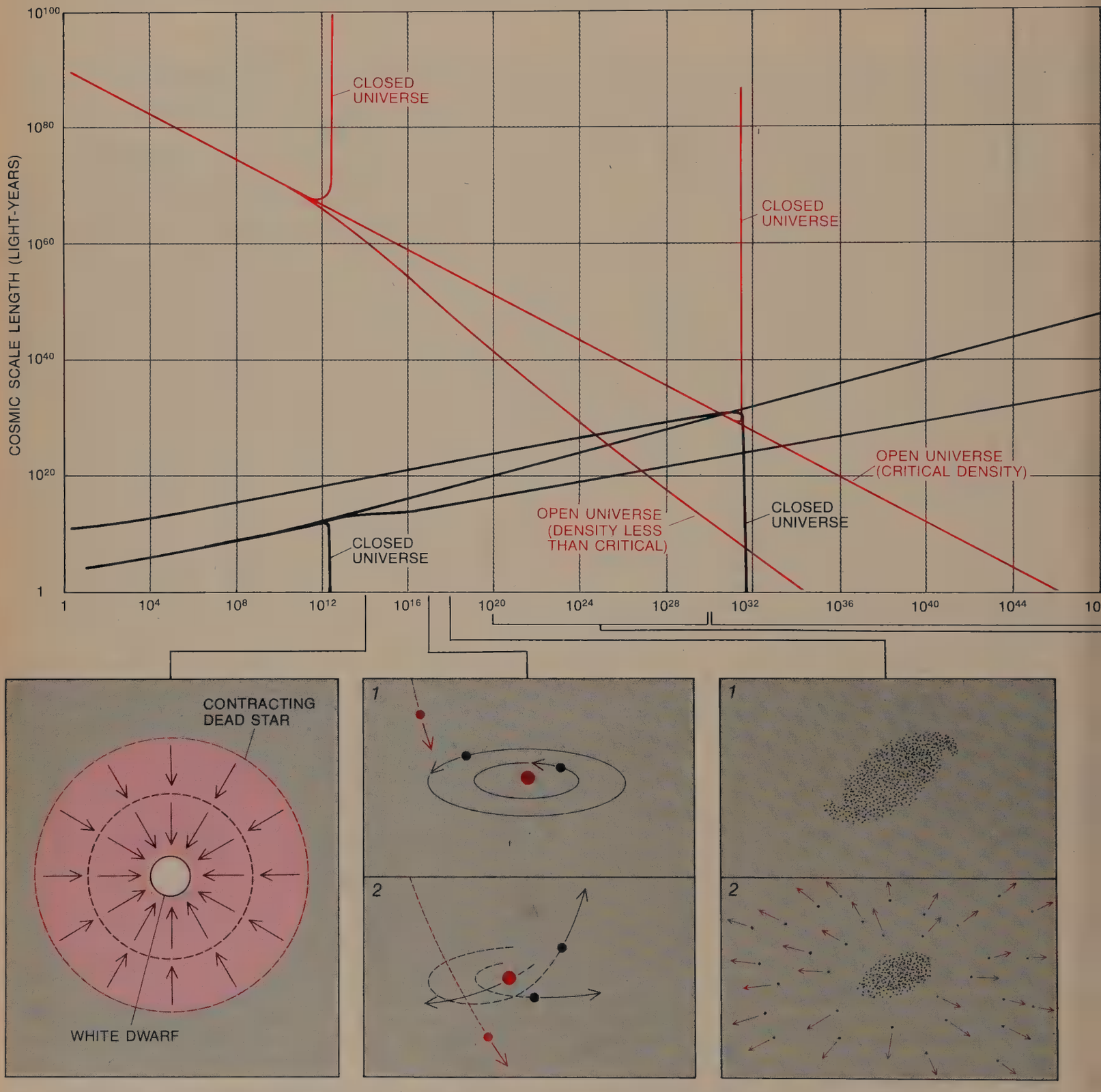
Similarly, the observed motions of galaxies in clusters of galaxies appear to require enough nonluminous matter to make up 50 percent of the critical density of the universe. The nonluminous matter would explain why the galax-



MOST DISTANT OBJECT KNOWN is a quasar designated PKS 2000–330; it has a red shift of 3.78, which implies it is receding from us at 92 percent of the speed of light. The enormous recessional velocity is probably an effect of the big bang; if it is, PKS 2000–330 is now more than 16 billion light-years from our galaxy. The photograph is a negative print; the quasar is the dark, starlike object in the exact center. Such distant objects enable cosmologists to infer some of the conditions that prevailed in the distant past and so to extrapolate certain processes into the remote future. Light now reaching the earth from the quasar must have been emitted shortly after the big bang.

ies remain bound in clusters instead of moving off in random directions. The dark matter could be in the form of cold rocks, burned-out stars, clouds of gas or even black holes.

If the additional matter necessary to make up the remaining 50 percent of the critical density actually exists, it may consist of ionized hydrogen atoms or of neutrinos. Ionized hydrogen should be detectable from X-ray telescopes in orbit around the earth. As for neutrinos, there should be about as many of them as there are photons, or quanta of electromagnetic energy, associated with the cosmic background radiation; this amounts to a few hundred neutrinos per cubic centimeter. Their contribution to the density of the universe depends on the mass of the neutrino. It was once supposed the neutrino is massless, but current experiments only place upper bounds on neutrino masses and do not rule out the possibility that neutrino masses are sufficient to close the universe. Moreover, certain recent theoretical arguments suggest that the neutrino has some mass, and several groups of experimenters are trying to detect and measure it. Such experiments could definitely answer the question of closure in the affirmative within the next few years. A negative result, on the other hand,

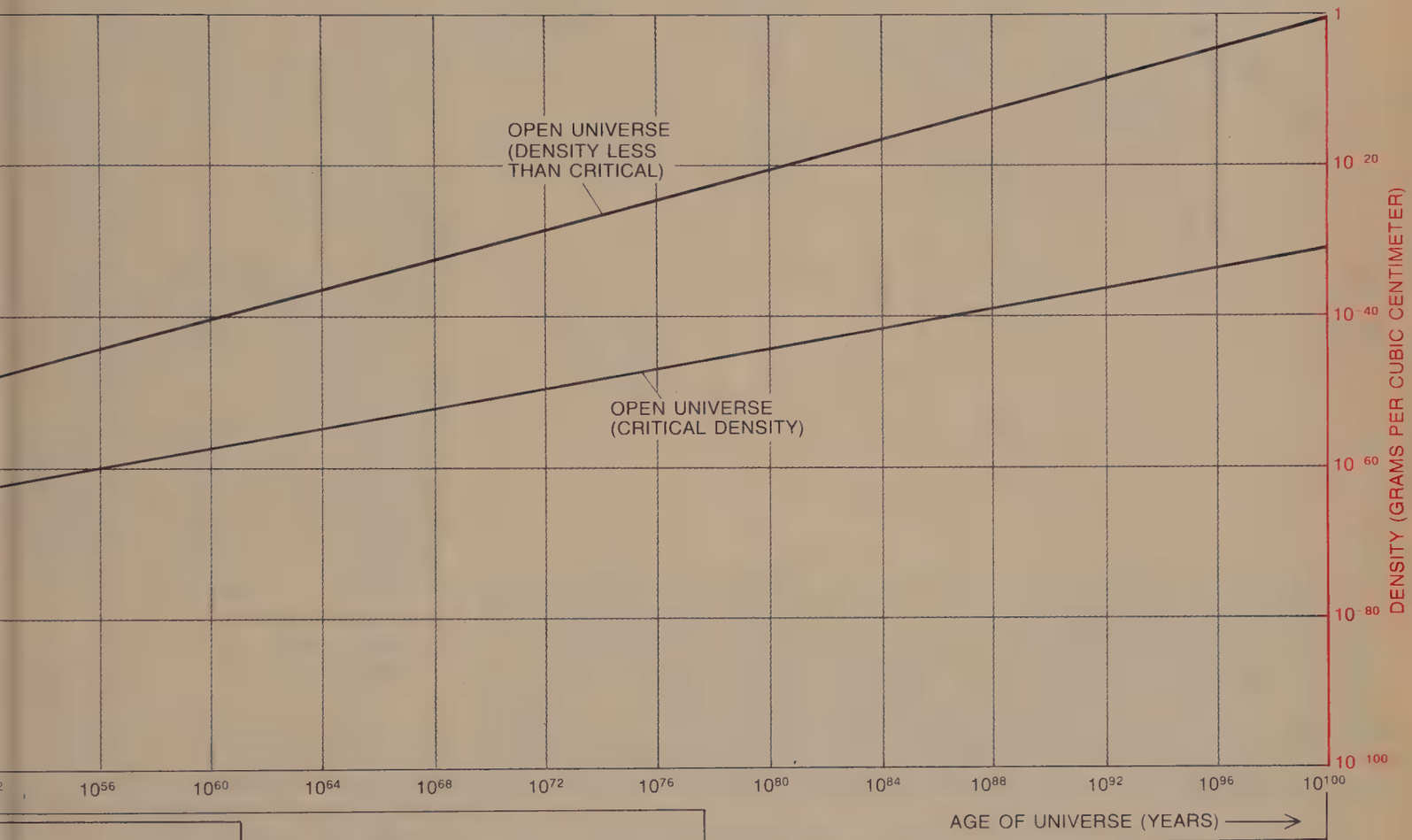


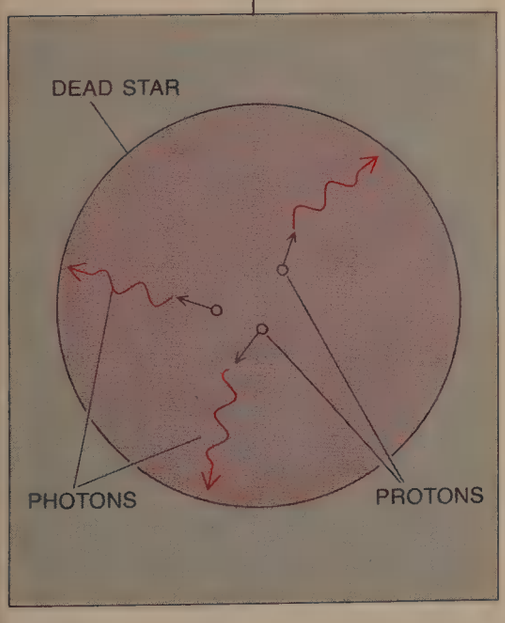
SIX MAJOR EVENTS in the future of the universe are expected on the basis of physical theory, provided the expansion of the universe initiated by the big bang continues long enough for them to take place. Here the scale length (which is roughly equivalent to the radius of the universe) and the density are plotted against time for four models of the universe: one in which the universe is open, two in which it is closed and one on the mathematically defined boundary between the closed and the fully open models. If the universe is open, the density of matter and energy is too low for gravity to overcome the expansion, and so the universe will expand forever. If the universe is closed, gravity will eventually halt the expansion and drag the universe back into a point of infinite density. It is not known whether we inhabit an open or a closed universe. The future events are illustrated schematically: the stars run out of fuel and contract

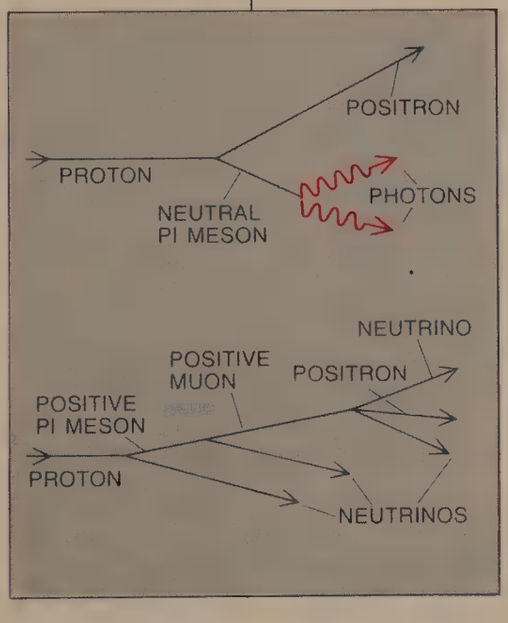
would leave the question to be resolved by more precise measurement of the distances and recessional velocities of remote galaxies.

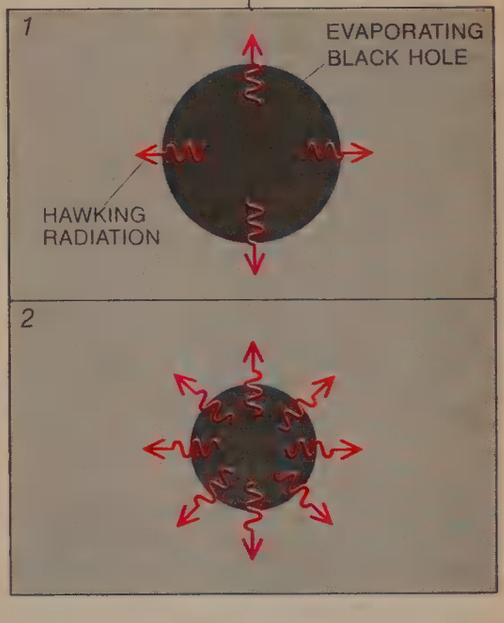
Although the current uncertainty about the density of the universe straddles the value that distinguishes a closed universe from an open one, the uncertainty is confined to less than two orders of magnitude. The fact that the universe has not contracted before reaching its present age implies that the density is not greater than about 10 times the critical density. The estimate of density based on luminous matter alone indicates that the actual density must be at least a twentieth of the critical density. No one yet knows how to account for what cosmologists find to be a striking coincidence.

Since the question of closure has not yet been settled on experimental or observational grounds, both the open and the closed possibilities need to be considered in any speculations about the distant future. Suppose, to begin, that the critical density is not attained and the universe is open. What is likely to happen? The question must be answered both for large-scale structure (that is, for the evolution of the geometric properties of the universe) and for local com-







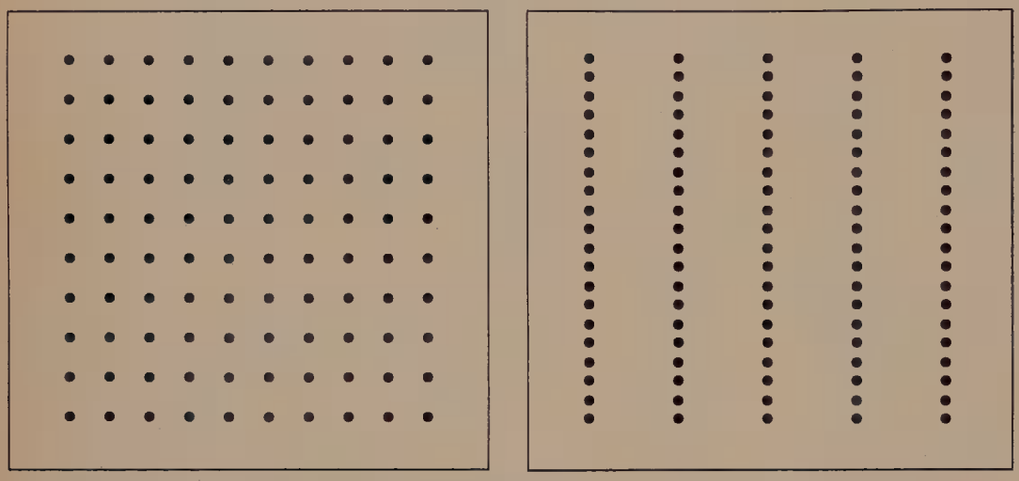


under their own weight in 10^{14} years; they lose their planets through close encounters with other stars in 10^{17} years. Galaxies lose a large fraction of their mass in about 10^{18} years because of still closer encounters between stars and because of scattering of both hydrogen atoms and dust particles; aggregates of matter that acquire a large velocity through the encounters can overcome the gravitational forces within the galaxy, and the remaining matter collapses into a supermassive black hole. (A black hole is a region of space from which nothing can escape except through mechanisms described by the quantum theory.) The decay protons generate heat inside the cold stars that escaped from the galaxies. The heating becomes significant in 10^{20} years, and by the year 10^{30} some 40 percent of the matter in the universe has undergone such decay. After 10^{100} years the supermassive black holes lose their mass through quantum "evaporation."

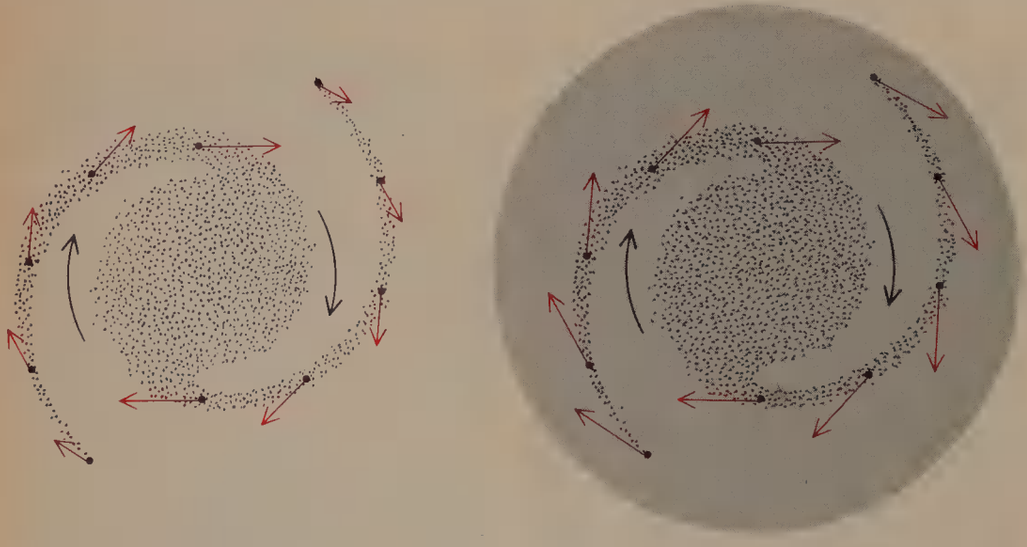
position (for aggregates of matter that range in size from protons to galaxies).

According to our best understanding, the local composition of the open universe will undergo six major transitions. First, within 10^{14} years after the big bang all stars will have run out of fuel. The principal nuclear fuel is hydrogen, which is converted into helium in the core of a star throughout most of its lifetime. After most of the hydrogen fuel has been consumed the star rapidly swells to many times its former size and becomes a red giant; in this stage of the star's existence helium is usually converted into carbon and heavier elements. These thermonuclear reactions are effectively one-way transitions: hydrogen is converted into helium, helium into carbon and carbon into heavier elements in a series that generally terminates with iron. The iron nucleus has a lower total energy per unit of mass than any other nucleus, so that once the "iron limit" has been reached the energy of the universe that is stored as nuclear fuel will have been entirely released.

The rate at which nuclear fuel is consumed by a star depends on the mass of the star: the more massive the star is, the faster it burns and so the shorter its lifetime is. The sun, for example, will use up much of its hydrogen within about 10 billion years, whereupon it will rapidly fluctuate in size, burn through some of the heavier elements at a prodigious rate but for a relatively short time and finally collapse into a small, slowly cooling white-dwarf star. The smallest stars will require many times the life span of the sun to pass through all these stages, but eventually they too will reach the iron limit. Note that although the exhaustion of nuclear fuel is the first major event in the future of the open universe, it will happen at a time exceedingly remote from our own. The last stars will cease to shine only after the universe is 10,000 times as old as it is now.



DISTRIBUTION OF MATTER AND ENERGY in the universe appears to be both homogeneous and isotropic to a high degree of precision. In a homogeneous universe equal numbers of galaxies fill comparable volumes of space. In an isotropic universe the density of matter and energy is the same in all directions as well as being homogeneous. The two properties are not equivalent: a homogeneous universe need not be isotropic. If matter is distributed as it is in the schematic diagram at the left, the universe is both homogeneous and isotropic. The distribution shown at the right is homogeneous but not isotropic. The distance between galaxies along the vertical direction is less than the distance along the horizontal direction. Anisotropy in an expanding, homogeneous universe is created by an expansion rate that varies with direction.



NONLUMINOUS MATTER may constitute a significant part of the total mass of the universe. There is evidence that such matter makes up at least half of the mass of most galaxies. If galaxies were made up entirely of matter that can be observed directly, stars at the periphery of a galaxy would revolve about the galactic center with a velocity that decreases with distance from the center (*left*). The observed velocity of peripheral stars, however, does not decrease with distance. The discrepancy can be explained by assuming that nonluminous matter is distributed throughout the galaxy and in a halo around it. Stars that appear to be on the periphery, therefore, would actually be in the midst of galactic matter, and the distribution of their velocities with distance from the galactic center would more closely match the observed distribution (*right*). The lengths of the arrows in the diagrams are proportional to the velocities of the stars.

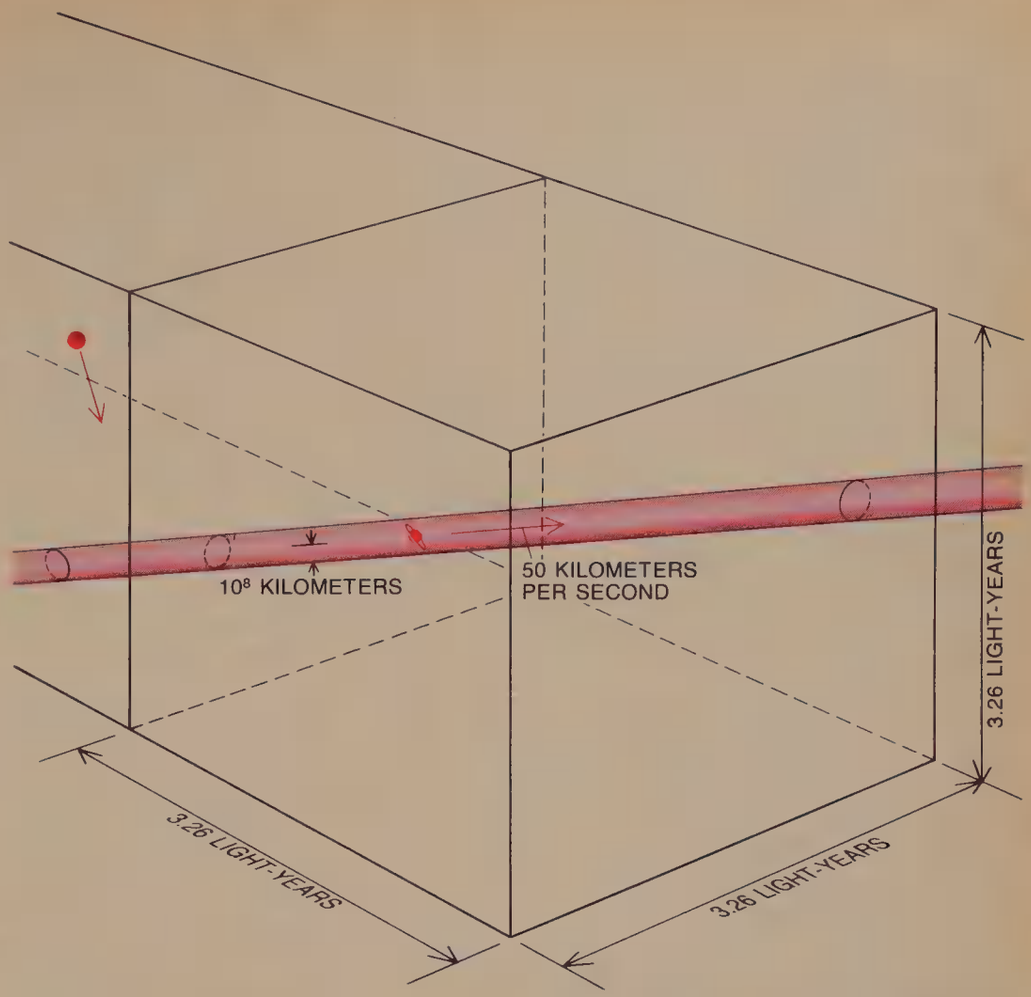
The second major event is that all stars will lose their planets. If a star that has a planet is approached by another star to within the radius of the planetary orbit, the orbit will be markedly disrupted by the gravitational field of the passing star, and the planet can be scattered into space. The average time before such an encounter can be expected depends on the density of stars in a given region, on the area of the planetary orbits and on the velocity of the stars with respect to one another. The density of stars can be expressed in terms of the volume in which at least one star is likely to be found. A star moving through space with a planet traces a cylindrical volume whose size depends on the area of the orbit of the planet and on the star's speed. The average interval between stellar encounters is equal to the time required for the volume of the cylinder to become as large as the volume in which a star can be expected [*see illustration on opposite page*].

The density of stars in a galaxy is one star per 35 cubic light-years. Freeman J. Dyson of the Institute for Advanced Study in Princeton suggests that a reasonable value for the radius of a planetary orbit is about 100 million kilometers, and he notes that a typical star pulls its planets through space at about 50 kilometers per second. The volume of the cylinder traced by the moving planetary system will therefore be 35 cubic light-years after about 10^{15} years, and so a disruptive encounter is likely in this interval. It can be safely assumed that after 100 such encounters all of a star's planets will have been kicked out of orbit, so that in about 100 times 10^{15} years, or 10^{17} years, all stars will have lost their planets.

The third transition foreseen is the result of still closer stellar encounters; their effects are evident on a galactic scale. When two stars come close to each other, the gravitational interaction can transfer kinetic energy from one star to the other. If the encounter is close enough, one star can gain so much ener-

gy that it attains the velocity needed to escape from the galaxy. Because energy is conserved in the interaction, the second star in the pair must lose kinetic energy, and as a result it becomes more tightly bound to the core of the galaxy.

The process can be called galactic evaporation, because the stellar interactions reproduce on a majestic scale the interactions of molecules evaporating at the surface of a liquid. A similar exchange of energy can also cause an appreciable fraction of the interstellar gas to escape from the galaxy. After as much as 90 percent of the mass in the galaxy has evaporated, the gravitational field will draw the remaining stars and dust into an increasingly dense core. Galaxies in their present form are likely to include a central, supermassive black hole, a region of space from which escape is impossible (except through mechanisms described by the quantum theory). If no such black hole exists now, the galactic core would probably still grow so dense that gravitation would overwhelm the resistance offered by gas pressure, and the core would collapse anyway to form a supermassive black hole. A calculation similar to the one we have outlined for the loss of planets shows that the evaporation of stars and the collapse of galaxies should be complete after about 10^{18} years.



FREQUENCY OF ENCOUNTERS between a star and a planetary system can be estimated from the number of stars per unit volume, the size of the planetary system and its velocity with respect to the stars. The encounters disturb a planetary orbit whenever the star intersects the volume of space filled by the orbit of the planet. On the average such an encounter will take place in the same time required for the planetary system to trace a cylinder in space equal in volume to the cube within which one expects to find a single star. There is one star in the galaxy for every 35 cubic light-years. Each planetary system moves at about 50 kilometers per second, and the outer planetary orbits trace a cylinder whose radius is about 100 million kilometers. Under these assumptions it is straightforward to calculate that one encounter takes place on the order of every 10^{15} years. After about 100 such encounters, which can be expected in 10^{17} years, the orbits of all planets will have been so seriously disrupted that the planets will have escaped from their parent star and scattered into interstellar space. The scale of the stars and the planetary system with respect to their density in space has been greatly exaggerated.

The fourth and fifth transitions we envision for the open universe are late cosmological effects predicted by most grand unified models of particle physics, but the effects do not become important until the universe is 100 times older than it is when galaxies collapse. The aim of the grand unified theories is to present a comprehensive account of the strong, the weak and the electromagnetic forces between elementary particles. At the relatively low energies accessible in terrestrial laboratories the three forces seem quite different; nevertheless, the theories describe them as manifestations of a single interaction whose unity can be perceived at the energy corresponding to a temperature on the order of 10^{27} degrees K. At that extraordinary temperature particles such as the quarks, which "feel" the strong force, can be transformed into particles such as the electron and the positron, which at lower energies feel only the weak and the electromagnetic forces. At lower temperatures such changes of identity become highly improbable, but they can still take place occasionally. The proton, which is thought to be a bound system of three quarks, could decay if its constituent quarks underwent such a transformation.

According to most grand unified theories, all protons should decay after a period of from 10^{30} to 10^{32} years; a decay rate of at least one proton per year should therefore be registered in any aggregate that includes 10^{32} protons, such as a mass of water weighing about 160 tons. There are now 13 experiments either in operation or under consideration that attempt to monitor large masses of water, iron or other materials for evidence of proton decay. Decay signals have tentatively been identified from 150 tons of iron being monitored 2,300 meters underground at the Kolar goldfields near Bangalore in southern India. A single candidate decay signal has also been registered by detectors installed in the Mont Blanc Tunnel between France and Italy. It is too soon to judge whether these signals will continue to stand up to scrutiny.

If the proton is susceptible to decay, the process will have a significant effect on stars that have not been captured by galactic black holes. Such stars are the ones that escape by evaporation, and the decay of the protons and neutrons in them will keep them considerably warmer than the surrounding interstellar medium. For an assumed proton lifetime of 10^{30} years the decay rate in a typical star the size of the sun is about 10^{27} per year. Each proton decay gives rise to a shower of energetic electrons, positrons, neutrinos and photons. All the daughter particles except the neutrinos are absorbed by the star, and the absorbed energy keeps the star warm.

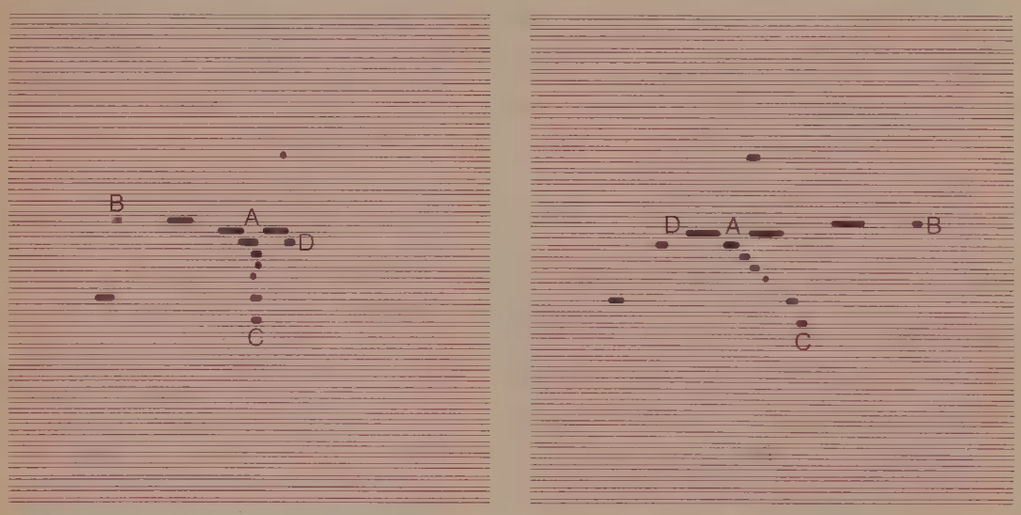
The precise temperature of the star during the era of proton decay can be determined by setting the rate at which energy is radiated from the star equal to the rate at which the decay releases heat energy. In this state of equilibrium the temperature depends on the mass of the star, on the surface area from which the heat can be radiated and on the rest energy and lifetime of the

proton. The equilibrium temperature is 100 degrees K. for the most massive dead stars (which, paradoxically, are the smallest ones) and about three degrees K. for larger and less massive ones.

The stars will cool to their equilibrium temperature in about 10^{20} years and thereafter their temperature will remain roughly constant until most protons have decayed, in about the year 10^{30}. The stellar glow is cold, but not in comparison with the temperature of the background radiation left over from the big bang. The background temperature depends on the details of the expansion of the open universe. If the density of the universe is less than the critical density, the background-radiation temperature will have fallen to 10^{-20} degree K. by the year 10^{30}. On the other hand, if the density is exactly equal to the critical density, the universe will have expanded more slowly and the temperature of the background radiation will have fallen to 10^{-13} degree K. That is, the temperature will be from 13 to 20 orders of magnitude less than the temperatures of the dead stars.

Proton decay also changes the constitution of interstellar gases that evaporate before galaxies collapse. Inside a star a positron released by the decay of a proton soon encounters an electron, and the two particles annihilate each other. The annihilation generates more photons and heats the star. In intergalactic space the density of matter is so low (and continuously decreasing because of the expansion of the universe) that positrons and electrons are extremely unlikely to collide. Indeed, by the year 10^{30} an open universe with a subcritical density will have expanded to more than 10^{20} times its present size, and the average distance between an electron and a positron in interstellar space will be of the same order of magnitude as the size of our galaxy. (If the density of the universe is equal to the critical density, the universe will have expanded by a factor of 10^{13}.) The interstellar medium therefore becomes an extremely rarefied gas, formed from about 1 percent of today's electrons and from positrons created by the decay of about 1 percent of today's protons.

The events associated with the decay of the proton will have played themselves out by the time the universe has existed for 10^{32} years, or 100 times as long as the average proton lifetime. What remains of the universe then is the rarefied electron-positron gas, the photons and neutrinos that were emitted in various earlier epochs and the supermassive black holes. The photons and neutrinos are left over from the big bang, from the years when the stars shone, from the decay of protons and neutrons throughout history and from the final decay of dead stars. Photons and neutrinos lose energy, and both they and the other constituents become more rarefied as the expansion continues. In other respects the universe at subcritical density remains quiescent until roughly the year 10^{100}, a period lasting 10^{68} times as long as all the processes we have described so far.

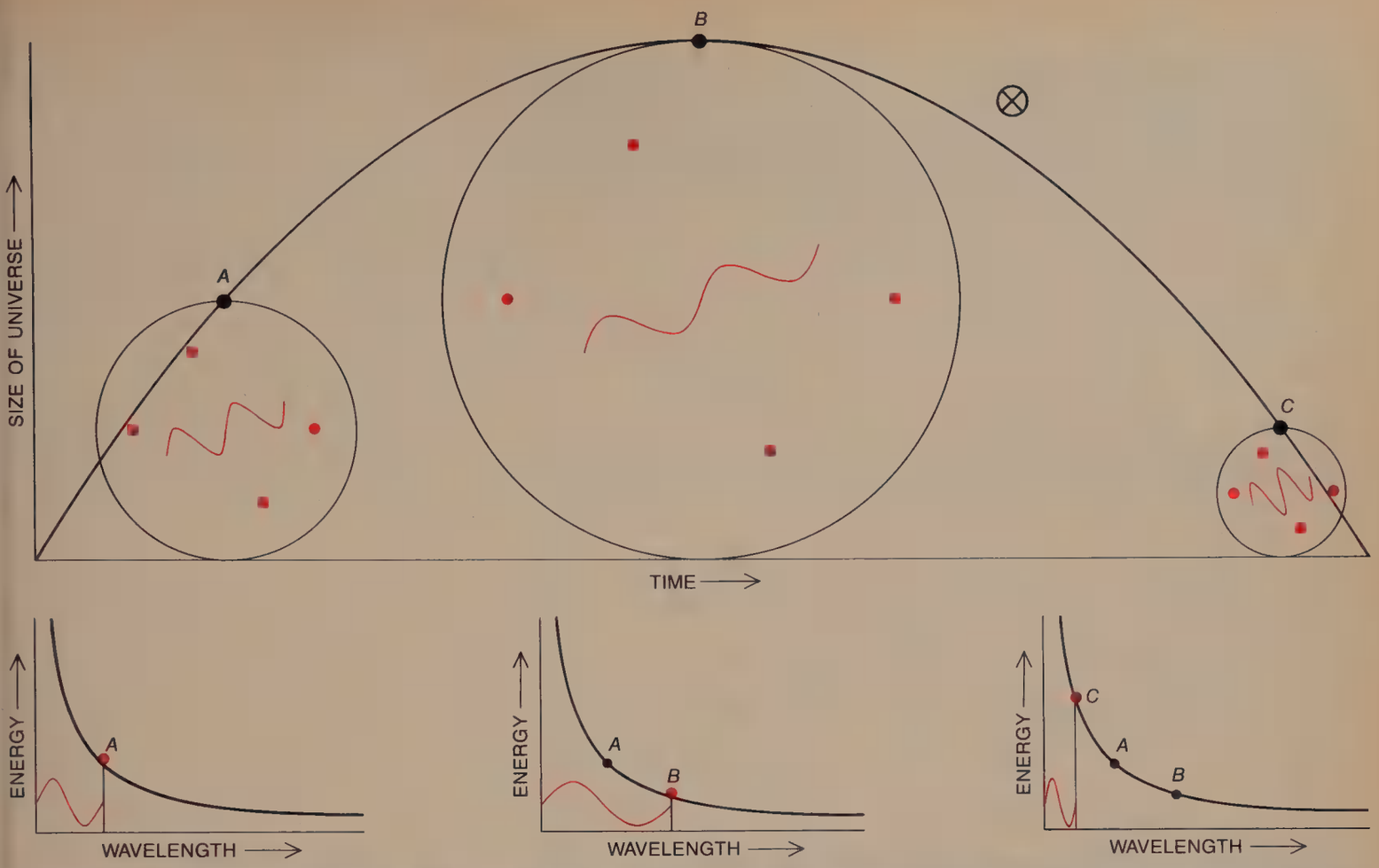


DECAY OF THE PROTON is a possible event that would have important consequences for the future of the universe. The decay should be observed experimentally if it is not too rare. Most of the current "grand unified theories" of elementary-particle interactions predict that the lifetime of the proton is between 10^{30} and 10^{32} years. Some 13 experiments now under way or in the planning stages are trying to detect the decay. Candidate decay events from two experiments are under study. Some of the events were detected by investigators working in a mine in the Kolar goldfields near Bangalore in southern India; another event was seen by investigators from the European Organization for Nuclear Research (CERN), working with a detector in the Mont Blanc Tunnel between France and Italy. The Mont Blanc event is shown in two schematic views at a right angle to each other. The parallel planes are plates of iron that are monitored for decays, and the black dots represent the products of decay. If proton decay can be verified, an evolutionary milestone can be fixed at about the year 10^{30}. All long-lived matter except electrons, positrons, photons, neutrinos and black holes will have decayed by then.

The sixth and last major event in the future of the open universe is the decay of the black holes. In the simplest interpretation of Einstein's gravitational theory nothing can escape from a black hole. There is a boundary called the event horizon where the velocity needed to escape is equal to the speed of light; no particle inside the event horizon can ever go fast enough to cross it. Nevertheless, in 1974 S. W. Hawking of the University of Cambridge showed that a quantum-mechanical phenomenon implies that a black hole can give up all the energy associated with its mass and thereby disappear.

The quantum phenomenon is a consequence of the uncertainty principle of Werner Heisenberg, which states that it is only with a certain probability that either the momentum or the position of any particle can be determined to be within a given range of values. More precisely, the product of the uncertainty in the momentum of the particle and the uncertainty in its position is not less than a numerical constant. In classical physics (that is, in physical theory that does not take account of quantum phenomena) a particle can cross an energy barrier only if it is given enough energy to surmount the barrier. The event horizon is an absolute energy barrier in classical physics: a particle could not gain enough energy to surmount it. Because of the uncertainty principle, however, a particle initially found in one region can later be found in another region, even if its classical energy is much less than the height of the energy barrier between the two regions. A particle that crosses an energy barrier in this way, without acquiring the energy necessary to surmount it, is said to have tunneled through the barrier. Hawking pointed out that because particles can tunnel through the event horizon, mass and energy can be ejected from a black hole [see "The Quantum Mechanics of Black Holes," by S. W. Hawking; SCIENTIFIC AMERICAN, January, 1977].

Hawking showed that the rate at which energy is emitted by the black hole is inversely proportional to the square of its mass. The rate is initially low, but as the mass decreases, the energy loss accelerates. It follows that all black holes must eventually disappear, or evaporate; for the supermassive black-hole relics of collapsed galaxies the evaporation time is about 10^{100} years. Most of the decay products are photons; the emissions become increasingly energetic during the later stages of the evaporation. After 10^{100} years, therefore, the universe is made up of an extremely diffuse gas of electrons, positrons and neutrinos, low-energy photons emitted long before black-hole decay and numerous expanding spheres of



ENERGIES OF FREE PARTICLES are not conserved in a closed universe. Because of cosmic expansion and contraction in such a universe, phenomena that depend on distance vary with time, just as the distance between two points on the surface of a balloon increases or decreases as the balloon is inflated or deflated. Thus the wavelength of a photon increases during expansion and decreases during contraction. The energy of the photon is proportional to the reciprocal of the wavelength, and so the energy of the emitted photon decreases during the expansion and increases during the contraction. When the wavelength of the photon is shorter than it was at the time of its emission, there is a net gain in the energy of the universe as a whole. The additional energy could prolong a new, subsequent expansion.

high-energy photons emitted by evaporating black holes.

Mathematical models of the large-scale structure of an open universe suggest that the universe we live in is a very special place. To describe the universe mathematically six quantities must be specified: three of them give the rate of expansion along the three directions in space and three give the rate of change in the expansion rate (that is, its acceleration or deceleration) along the three directions. If the six quantities are specified at some definite time and if the distribution of energy is given for the same time, the values of the six quantities and the distribution of energy are determined for all later times.

In 1973 Hawking and C. B. Collins, also of Cambridge, showed that for almost all possible values of these quantities at an early stage in the big bang the universe must become increasingly anisotropic as it expands. The average separation of galaxies along one spatial axis ought to become increasingly different from the average separation along the other axes because of differences in the expansion rates. The result suggests that the actual universe is a highly improbable one. For example, as we noted above, the cosmic microwave background is invariant with direction to within one part in 10,000. On a cosmic scale the matter in the universe is also smoothly distributed. A major unsolved problem in cosmology is to understand why the universe is so nearly isotropic.

One approach to the problem is to choose initial values of the six quantities in such a way that the energy density of matter and radiation is exactly equal to the critical density. If such a universe is dominated by matter instead of radiation throughout its history, any initial anisotropy will not grow. The decay of the proton, however, can lead to a radiation-dominated universe.

John D. Barrow and Frank J. Tipler of the University of California at Berkeley have shown that in a universe at the critical density the electrons and positrons remaining after proton decay will begin to form bound pairs after more than 10^{70} years. Each bound pair is effectively an atom, but the electron and positron orbit each other in such a highly excited state that the region of space encompassed by the pair is larger than the present observable universe. The encompassed space grows progressively larger the later a bound pair is formed.

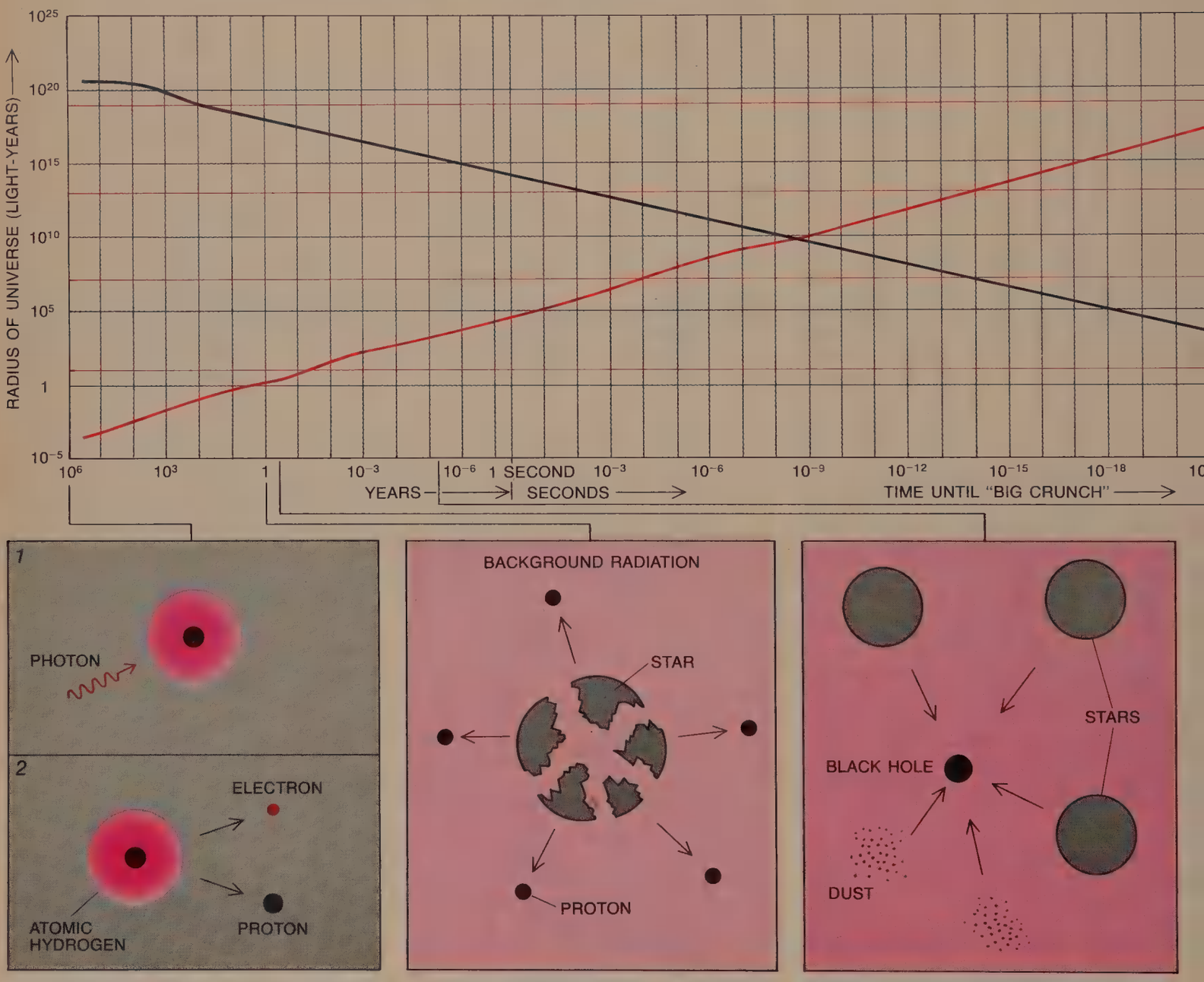
As time goes on the electron and the positron spiral inward and eventually annihilate each other, giving rise to high-energy photons. The wavelength of the photons is shifted toward the red end of the spectrum as the universe continues to expand. In the model described by Barrow and Tipler the universe finally comes to be dominated by radiation and the anisotropy increases. In a subsequent investigation, however, Don N. Page and M. Randall McKee of Pennsylvania State University showed that under certain conditions the electron-positron pairs can continue to dominate the energy of the universe even though their number is decreasing. The red shift of the photon radiation decreases its energy and the increasingly long time required for the annihilation of late-forming bound pairs causes a decrease in the rate of photon emission. Hence the anisotropy that is characteristic of a radiation-dominated universe may not be able to grow.

All the foregoing speculations concern an open universe. We shall now consider the future of the universe on the assumption that there is enough nonluminous matter for gravity to halt the

expansion and bring about a contraction. The closer the average density is to the critical density, the longer the closed universe will last. We know of no reason, however, for the average density to be close enough to the critical density for the universe to survive until most protons decay. The universe at its maximum expansion is therefore likely to be made up of dead stars, supermassive black-hole remnants of collapsed galaxies and low-energy photons and neutrinos, just as it would be if the universe were open.

One curious aspect of a closed universe is that although energy is conserved locally, the total mass or energy of the universe is not conserved. For a given size of the universe the total energy during contraction is greater than it is during expansion. Consider a photon emitted from the sun toward intergalactic space. The conservation law is satisfied during the emission because the energy carried by the photon is exactly balanced by a very small decrease in the mass of the sun. As the universe expands, the wavelength of the photon increases in proportion, with the result that its energy diminishes. When the universe contracts, the photon's wavelength contracts as well and its energy increases. Eventually the wavelength of the photon becomes shorter than it was at the time of emission, so that the photon has gained energy without any compensatory loss of mass or energy in some other region. The universe therefore runs hotter in contraction than it does in expansion. The photons that make the largest contribution to the extra energy are the ones emitted when the universe is at or near its maximum size, because such photons will undergo the most contraction.

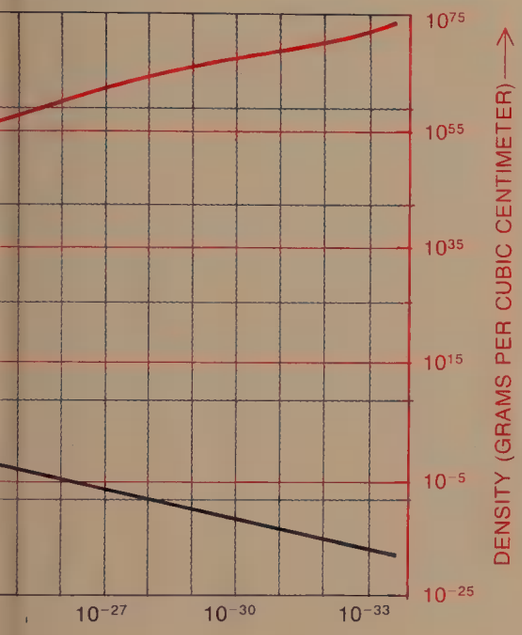
The major events during the expansion phase of a closed universe follow the same sequence as the events in an open universe. (There is, of course, no collapse for an open universe.) Several investigators have studied the collapse, including Martin J. Rees of Cambridge. As the photons gain energy during the contraction they heat the dead stars, causing them to begin to burn rapidly, explode or evaporate. The resulting

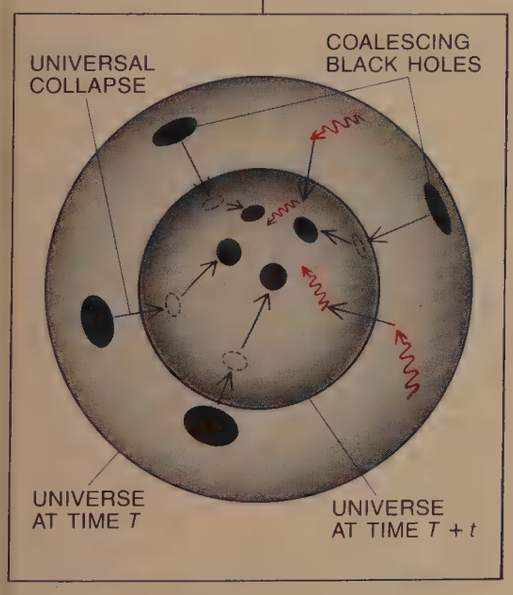


FINAL STAGES in the collapse of a closed universe would very roughly reverse the stages in its expansion if it were not for the effects of black holes. About 20 billion years before the big crunch, or the complete gravitational collapse of a closed universe, the universe will contract to a volume at which its energy density will be the same as it is now. As the contraction continues photons will become more energetic and the universe will heat up; about a million years before the big crunch the photons will dissociate interstellar hydrogen atoms into their constituent electrons and protons. A year before the crunch the temperature outside stars will become greater than the temperature inside, and the stars will begin to break up. At about the same time supermassive black holes will begin to swallow up the remaining stellar material as well as other matter and radiation. About three minutes before the crunch the supermassive black holes will be-

soup of particles would continue to retrace the steps in the expansion of the universe if it were not for the effects of the black holes. As the density increases, the black holes gobble up material, and they coalesce whenever they collide. One can calculate that in a universe with one supermassive black hole per galaxy the dead stars are swallowed by black holes shortly after the stars break up and material begins to evaporate from them. All the black holes finally coalesce into one large black hole that is coextensive with the universe.

Theoretical physics in its current state of development cannot fully describe the collapse of a black hole; it is not yet known how to extrapolate the





gin to coalesce; the coalescence is represented on the contracting surface of a sphere, which corresponds to the contraction of space. The variation in the density and radius of the universe is graphed with time; it was calculated numerically by applying Einstein's equations and other principles described in the text.

THE VISUAL DISPLAY OF QUANTITATIVE INFORMATION

EDWARD R. TUFTE
Yale University

"A landmark book, a wonderful book"
Frederick Mosteller
Harvard University

"A tour de force"
John W. Tukey
Bell Labs and Princeton University

Theory and practice in the design of statistical graphics.

Graphical excellence, with 75 examples of the finest data graphical work from 1700 to 1982. Time-series, thematic maps, relational graphics, multivariate designs, high density graphics.

Graphical integrity and sophistication. History of graphical deception. 40 graphical lies. Design variation vs. data variation. Causes of good and bad design. The bureaucratic setting of graphical production.

The theory of data graphics, including many new graphical designs. Chapters on data-ink and graphical editing, chartjunk, data-ink maximization and the derivation of new designs, multifunctioning design elements, data density, and aesthetics and technique in data graphical design. Graphical architecture and effective multivariate designs. Use of color. Computer graphics. Measures of graphical performance. Sentences, tables, and graphics in communicating quantitative information.

250 illustrations, including many of the most sophisticated and beautiful statistical graphics ever drawn. Color, clothbound, printed by The Meriden Gravure Company.

$32 postpaid. Order directly (enclosing check or money order) from Graphics Press, Box 430, Cheshire, Connecticut 06410

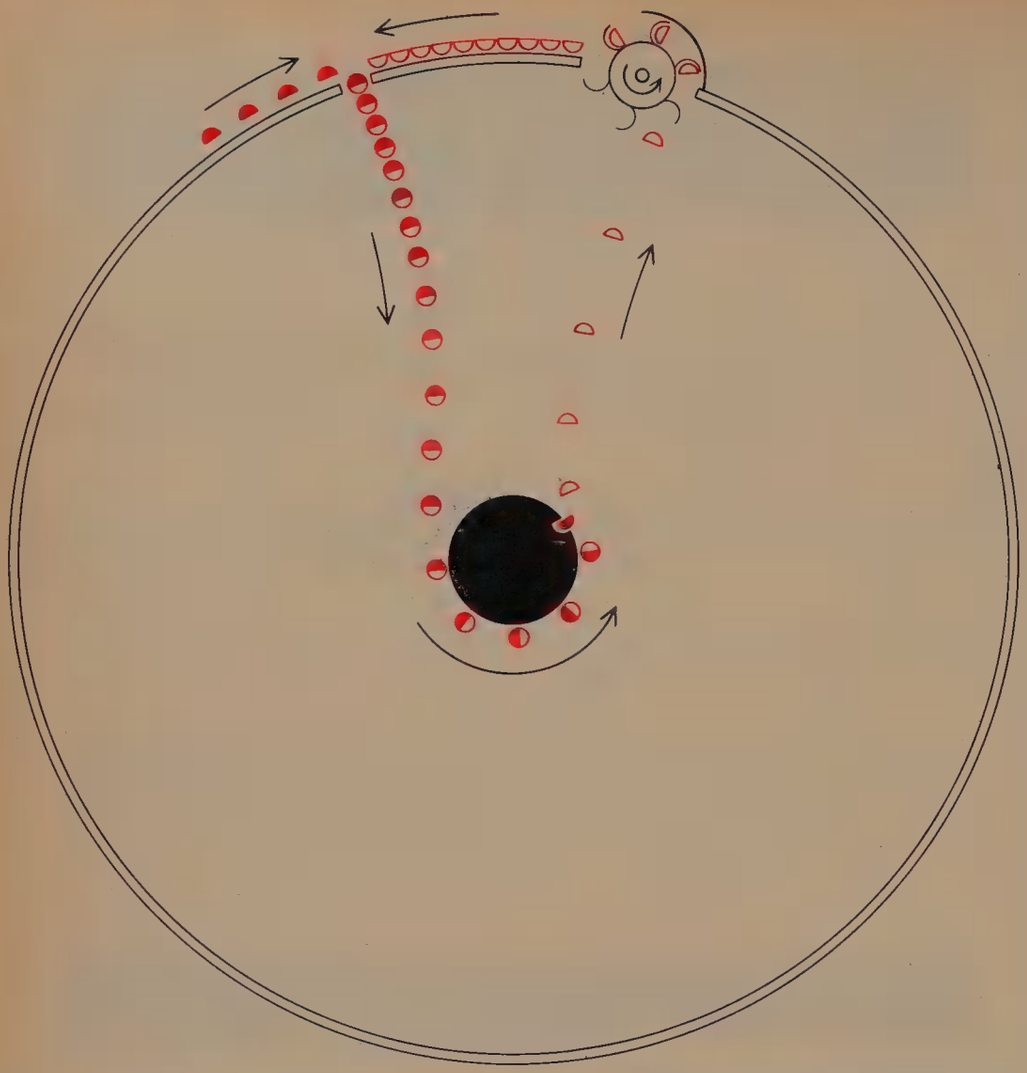
In Milan an appointment with science
SECOND INTERNATIONAL SCIENTIFIC & TECHNICAL BOOK FAIR
MILAN FAIR GROUNDS
14 - 18 APRIL 1983
S&T
For information:
c/o PROVINCIA DI MILANO
ASSESSORATO ALLA CULTURA
Via Guicciardini, 6
20129 MILANO (Italy)
Tel. 02/77402092 - 77402477

Free!
Edmund Scientific Catalog

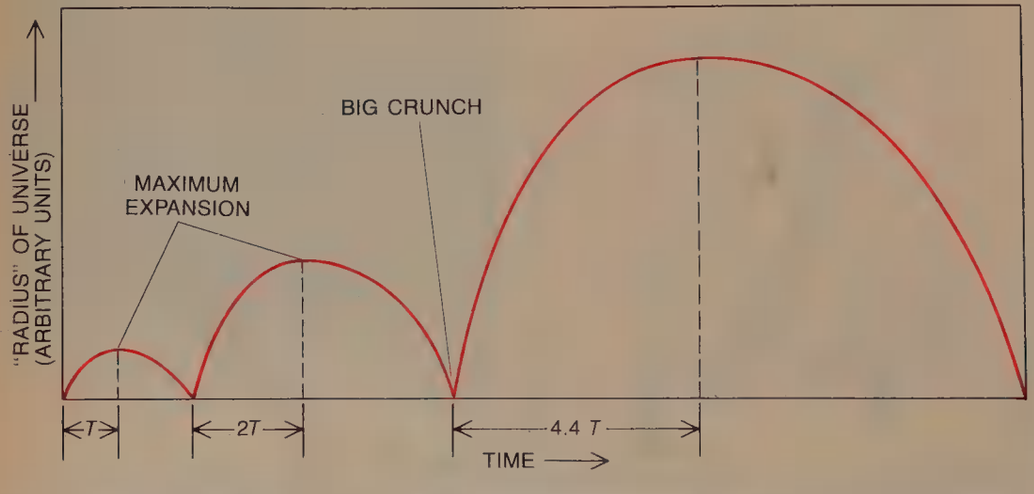
4,000 Unique Items

Astronomy, Microscopy, Bioteedback, Weather, Alternate Energy, Binoculars, Optics, Magnets, Magnifiers, Tools, Unique Lighting, Lab Equipment, and much more. Over 4,000 unique and fascinating products. Send for our FREE, colorful Edmund Scientific Catalog. Today!

Rush me your free catalog!
Name _____
Department _____
Institution _____
Address _____
City _____
State _____ Zip _____
Clip And Mail Coupon Today To:
Edmund Scientific, Dept. 8333 2110
Edscorp Bldg., Barrington, N.J. 08007



COSMIC ENERGY CRISIS could be delayed for a long time by the existence of black holes formed from collapsing galaxies. In principle energy can be extracted from the gravitational field of a black hole; hence it is conceivable that an advanced civilization could maintain itself for at least 10^{100} years, or many times as long as it takes for black holes to evaporate. The most efficient known scheme for extracting the energy of a black hole was put forward by Charles W. Misner of the University of Maryland at College Park, Kip S. Thorne of the California Institute of Technology and John Archibald Wheeler of the University of Texas at Austin. A rigid shell is constructed around a rotating black hole, and surplus material from the civilization is sent by a satellite close to the event horizon, the boundary of the region from which nothing can escape. Near the event horizon the material is jettisoned at the equator of the black hole. As a result both the gravitational potential energy of the jettisoned material and the rotational energy of the black hole are decreased. Because the total energy of the satellite and the black hole must be conserved, the energy of the satellite is increased by an amount equal to the two energy reductions. In principle the energy gained by the satellite could be converted into a useful form.



UNIVERSE MAY BE CYCLIC if the density of matter is sufficient to close it and if physical processes not yet unknown allow the universe to "bounce," or expand once again. Because energy is gained during the contraction phase, each new cycle expands for a longer time than the preceding one. The period of expansion should increase by a factor of about two during the next cycle of the universe, assuming that the current expansion continues for about 100 times the present age of the universe. The duration of succeeding cycles increases by larger factors.

equations that govern the large-scale structure of the universe back to a state of infinite density. It is possible, however, that before the density becomes infinite an unknown mechanism could cause the universe to "bounce" and begin expanding once again.

If the universe can bounce, it could remain closed and so be cyclic. The energy gained by the photons during each period of contraction might be conserved through the bounce; with each successive cycle the universe would therefore be larger, for a given temperature, than it was in the previous cycle and would also take longer to reach its maximum size. Robert H. Dicke and P. J. E. Peebles of Princeton University and the four of us have calculated that, if our universe is cyclic, the next cycle should expand for about twice as long as the expansion phase of the current cycle. For earlier cycles the expansion factor would be smaller; we calculate that the present universe is at most 100 bounces from the cycle that lasted just long enough to make a single generation of stars. The major difficulty for the bounce model is to understand how an extraordinarily inhomogeneous and locally anisotropic universe of isolated, coalescing black holes can be smoothed out; for that matter, there is no understanding of how a bounce can take place.

If the universe is cyclic, and if reactions mediated by the high temperatures available at the beginning of each cycle give rise to the excess of nuclear particles over antiparticles that is observed today, one can draw a strong conclusion. The evolution of the universe from the beginning of each bounce until the formation and dissolution of galaxies will follow the same course of events during each cycle. As the cycles become longer, however, the phenomena we have described for the later stages of the open universe will become important. For example, on the first cycle in which the expansion stage lasts longer than a ten-billionth of the average proton lifetime (that is, longer than about 10^{20} years) the energy gained during the contraction stage will be dominated by the photons emitted in the course of proton decay. Subsequent cycles may then lengthen by a factor of 1,000 instead of a factor of about two. Even larger expansion factors, on the order of 10^{12}, might result from the effects of coalescing black holes.

From a human point of view perhaps the most important question one can ask about the universe concerns the future of life and intelligence. Can the habit of thinking persist indefinitely in a universe increasingly hostile to life as it exists on the earth? Several cosmologists, including Dyson and Steven C. Frautschi of the California Institute of Technology, have begun to speculate

about issues such as the energy requirements for the indefinite maintenance of life and for communication among members of a society dispersed across increasingly large regions of space.

Dyson assumes that the embodiment of life and consciousness need not be limited to cells and DNA; instead, he maintains, the essential feature of consciousness is complexity of structure, which can be realized in any convenient material. Thus he assumes that the idea of a sentient computer or a sentient cloud cannot simply be dismissed as philosophically incoherent.

Given these assumptions, the changes in the environment caused by the death and cooling of stars and their evaporation from galaxies need not be insurmountable to a system that could be counted as alive and intelligent. For example, in principle energy could be extracted from the gravitational field of a supermassive black hole. The decay of protons and neutrons, however, may bring about a fundamental change, since it seems unlikely (although perhaps not impossible) that intelligence could be based on the tenuous foundation of electrons and positrons. Moreover, if the universe is closed, the conditions necessary for life may exist only for certain periods during each cycle.

In an open universe the ultimate impediment to life is quite different. With the evaporation of black holes there will be a cosmic energy shortage, because as the expansion of the universe continues the remaining particles and photons lose energy. Any constant rate of energy consumption by life forms will in due course become untenable. On the other hand, Dyson proposes that increasingly long periods of hibernation, during which energy is not consumed, could be interspersed with periods of consumption. Hence there is the potential for extraordinarily long-lived civilizations in this universe.

The starting point of contemporary cosmology, as we have emphasized, is the big-bang theory. Our quantitative treatment of the evolution of the big bang into the remote future is based on the new theories that describe the physics of elementary particles. In spirit, however, these ideas were expressed by the 11th-century Persian astronomer and poet Omar Khayyám, whose poem *The Rubáiyát* was translated by the English writer Edward FitzGerald. In one quatrain Omar wrote:

> With Earth's first Clay They did
> the Last Man's knead,
> And then of the Last Harvest sow'd
> the Seed:
> Yea, the first Morning of Creation
> wrote
> What the Last Dawn of Reckoning
> shall read.

HOLOGRAPHY HANDBOOK

Making Holograms the Easy Way

By Unterseher, Hansen and Schlesinger

Library Journal says:
"This comprehensive book is a guide for nontechnical people who wish to create their own holograms. Although holograms involve complex scientific principles, the authors describe and illustrate practical methods suitable for a 'handy' person at a relatively modest cost. Explicit information is given on equipment required and where to get it, how to construct optical tables, and how to make and display various types of holograms. A clear style and many explanatory diagrams and illustrations make the projects easy to follow. For those interested in theory the authors provide a section with a commonsense approach. Another section relating holography to the human brain and to cosmology will intrigue many who have an interest beyond the mechanics of holograms. Recommended - Frank Dayldoff."

THIS IS THE FIRST BOOK EVER PUBLISHED WITH A DAYLIGHT VIEWABLE HOLOGRAM IN EVERY COPY.

With only this book to guide them, the average person can go into their basement or garage and, for about what it would cost to buy a decent camera and set up a photographic darkroom, they can build their own holography lab and produce high quality holograms. This cost includes absolutely everything necessary including the laser.

This book is 8½" x 11", 408 pages and has hundreds of drawings along with 75 photos. Enclose check or money order for $19.50 (includes postage and handling). Foreign orders enclose $21.50 for sea mail and $29.75 if you wish air mail delivery. Address all correspondence to:

ROSS BOOKS
P.O. Box 4340
Berkeley, Calif. 94704

Play the finest

HARPSICHORDS & FORTEPIANOS

Build it yourself from one of our kits or let our experienced craftsmen build it for you.

Three Centuries of Harpsichord Making by Frank Hubbard (Harvard Univ. Press. 397 pgs. $20.00 postpaid)

For brochure send $1.00 to:
FRANK HUBBARD INCORPORATED
144-S Moody St., Waltham, MA 02154
(617) 894-3238

Free Catalog of Quality Outdoor Gear and Clothing

REI offers one of the largest selections of outdoor equipment, apparel and accessories. Whether you hike, climb, run, ski, kayak or cycle, REI has a unique assortment of goods, competitively priced, to send you on any pursuit.

Send for a free color catalog today and explore the world of outdoor adventure at REI. We guarantee satisfaction or your money back.

☐ Please send me a free catalog.

Name _____

Address _____

City _____

State _____ ZIP _____

REI
P.O. Box C-88127
Seattle, WA 98188
83SAM3
Quality Outdoor Gear and Clothing Since 1938

Memory in Food-hoarding Birds

Birds that hide seeds and later recover them appear to excel in spatial memory. One species evidently remembers where it has put thousands of caches for as long as several months

by Sara J. Shettleworth

In England in the winter a bird feeder stocked with peanuts is all the apparatus one needs to get an animal to show behavior suggesting it has remarkable powers of memory. Indeed, its memory appears to far surpass in capacity the memory that any other animal has ever shown in the laboratory. The feeder attracts many birds. Among them are great tits, blue tits and marsh tits, which are all small, lively birds related to the North American chickadees. The great tits and the blue tits congregate at the feeder, eating as fast as they can. They interrupt their meal only to chase away their competitors. A marsh tit nonetheless darts in, grabs a peanut and flies off. It is back almost immediately to grab another. It stores the peanuts nearby, each in a different site, until the feeder is empty. Then it searches out its hidden food.

In the American Southwest another bird, Clark's nutcracker, a relative of the jays and crows, shows similar behavior. In the late summer the nutcracker harvests the seeds of piñon pines. It repeatedly fills its sublingual pouch, a pocket under the tongue, and then flies several miles to bury the seeds. Often the burial sites are on bare, south-facing slopes where the snow will not be deep later on. A nutcracker may bury as many as 33,000 piñon-pine seeds in caches of four or five seeds each. Throughout the winter it returns and digs up its thousands of caches.

How do these birds find their hoards? Does the tit or the nutcracker remember where it has stored each peanut or cache of piñon seeds? Until recently observers of hoarding behavior in the wild doubted that the birds rely on memory. For one thing the bird would need a capacious memory indeed to remember the sites of hundreds or thousands of individually hoarded items. The memory would also have to be long-lasting. Even a short-term hoarder such as the marsh tit does not recover its stores until hours or days after it deposits them, and a long-term hoarder such as Clark's nutcracker does not return to a hoard for months, perhaps not until spring, when nutcrackers feed their young on piñon seeds. Moreover, in principle the hoards could be recovered without the aid of memory. A hoarder could store food only in certain kinds of sites, such as south-facing slopes or holes in bark. To recover its stores it would need only to search in sites of that kind. Then too it could employ memory only for the area in which it hoarded rather than for the individual sites. Inside that area it could search by trial and error or conceivably by cues such as odor.

Psychologists studying learning and memory in animals are becoming increasingly aware, however, that certain species have adaptive specializations that make them particularly good at learning and remembering things it is important for them to know. Among the well-known instances are the ability of many birds to learn their species' songs, the ability of rats and other animals to remember spatial locations, the ability of bees to remember the location of flowers and the ability of many animals to learn to avoid noxious food. If food-storing birds really do remember large numbers of storage sites over long periods, their memory could be another example of an adaptive specialization, one that would enable the birds to recover their stores efficiently. A bird that could remember where it had hidden its food would make fewer errors recovering it and spend less time and energy than birds searching randomly. Recent studies indicate that at least some food-storing birds do remember the sites of their hoards quite well.

Some of the most detailed evidence that food-storing birds remember their hoards has been accumulated over the past few years by John R. Krebs and his associates at the University of Oxford. They began by looking for evidence confirming that marsh tits in their natural environment recover their hoards. To that end marsh tits were trained to come to sunflower-seed dispensers set up in their territories in Wytham Wood near Oxford. Then Richard J. Cowie, Krebs and David F. Sherry coated the husks of the seeds with a radioactive substance harmless to the birds. (The birds remove the husks before eating the seeds.) The radioactivity enabled the investigators, equipped with a portable scintillation counter, to find the seeds the birds had cached and then check every few hours to see when they disappeared.

The seeds of course might be eaten by other birds or by rodents such as mice and voles. Cowie, Krebs and Sherry reasoned, however, that if marsh tits use their memory to recover their own stores, the seeds should disappear sooner from the birds' storage sites than they would from false sites the investigators established. This was indeed the case. The seeds the birds stored were depleted sooner than seeds the investigators placed in similar sites that were each one meter from a hoard made by a bird. In Wytham Wood in the winter the natural hoards were generally gone in a day or two. The last of the false hoards disappeared fairly soon afterward, suggesting that predation of the hoards is common.

One defensive tactic the marsh tits apparently adopt against animals that steal from their hoards is to disperse seeds rather than burying them in clusters. On the average the hoards the marsh tits established in Wytham Wood were seven meters apart. Sherry, Mark Avery and Allen Stevens set up false hoards of sunflower seeds and found that the seeds stayed in place longer the farther apart they were, up to a distance of about seven meters. Hoards farther apart than that were no safer than the ones at seven meters. In Wytham Wood, at least, marsh tits seem to have hit on the optimal spacing.

Having obtained evidence that marsh tits in nature recover their hoards, Krebs and his colleagues turned to observations in the laboratory. Marsh tits supplied with a bowl of seeds and with suitable storage sites hoard quite readily in captivity. Sherry, Krebs and Cowie offered marsh tits trays of moss one meter

on a side in which to store sunflower seeds. The birds did appear to remember where they had buried seeds even after 24 hours. In these experiments, however, the behavior of the birds was recorded only as visits to relatively large areas on the tray. Moreover, in an effort to ensure that the birds employed only memory in the experiments, the hoarded seeds themselves were removed from the tray before each bird was allowed to return. Thus it was difficult to say with what precision the birds located their hoarding sites. It was equally difficult to identify mistakes the birds might be making.

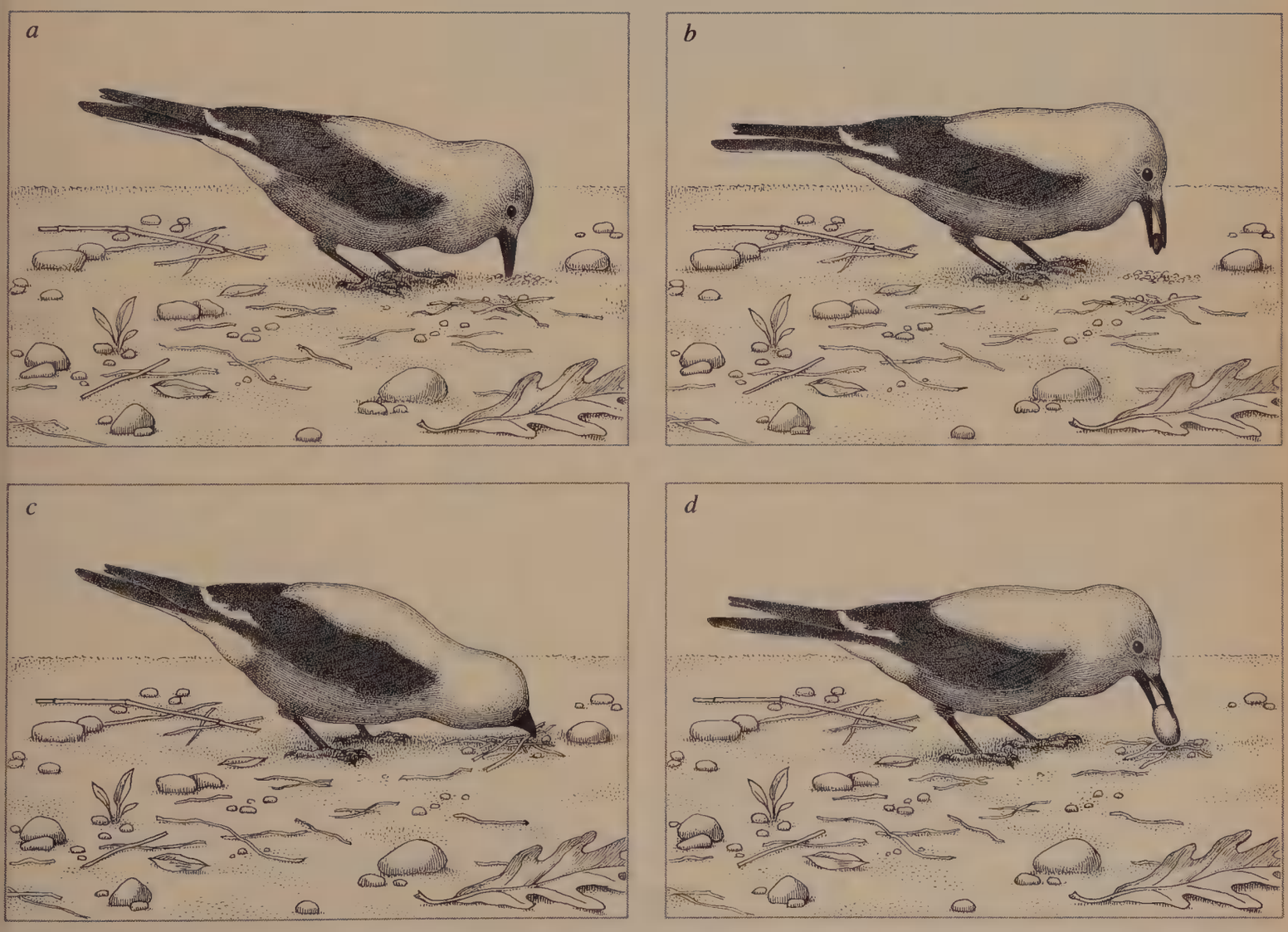
When a Guggenheim Fellowship enabled me to spend a sabbatical leave at Oxford, my interest in adaptive specializations of learning and memory attracted me to the marsh-tit studies. Krebs and I designed some experiments to analyze in more detail the marsh tits' apparent memory. We were particularly interested in finding out whether the items most recently stored by the birds are the ones they recover first. Such a "recency effect" often appears in tests of memory, and in the case of food hoarding there is good reason to expect it: if hoards are lost to scavengers at a constant rate, the most recent hoards will be the ones most likely to be available to the hoarders at any given time. Looking for them first is therefore the hoarders' best bet.

In a large room Krebs and I set up sections of tree branches. In each section we drilled a number of holes the right size for storing hemp seeds. There were about 100 holes in all. Each hole was covered with a flap of cloth, which a bird would have to lift in order to store a seed or look for one.

The room was the site of a series of experiments. For each trial of our first experiment we allowed a marsh tit to store 12 seeds it got from a bowl on the floor in the middle of the room. Each hole could hold only one seed; thus each seed was stored in a different hole. Once the seeds were cached we kept the bird outside the room for about two and a half hours. We removed the bowl of seeds; then we readmitted the bird, and it searched for the seeds it had stored. If it searched at random among the 100 holes, it would have to investigate an average of about eight holes to find one seed. The birds were much more efficient. On the average each bird made about two errors per seed. At the beginning of a recovery test a bird sometimes went to three or four seeds in succession without looking in empty holes.

Maybe instead of employing memory the tits were smelling the seeds or detecting them in some way we had not thought of. Smell seemed unlikely; the sense of smell is generally poor in birds. Still, to test these possibilities we allowed a marsh tit to store 12 seeds and then we moved the seeds to other holes (ones the bird itself had used in earlier trials). Under these circumstances the readmitted birds made about six errors per seed. They eventually found about two-thirds of the seeds by looking in more holes than usual.

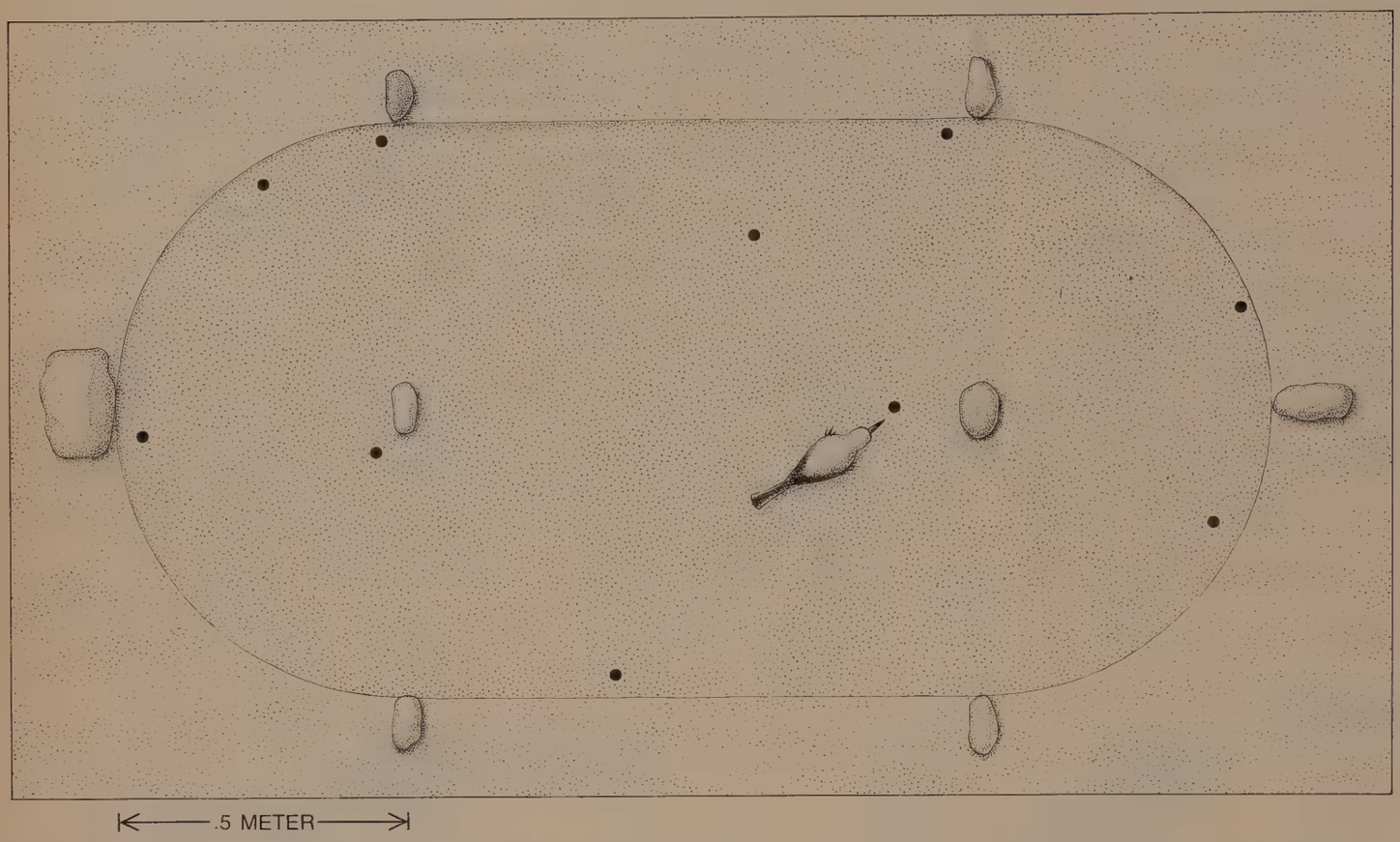
The results indicate that the marsh tits did not detect the stored seeds by smell

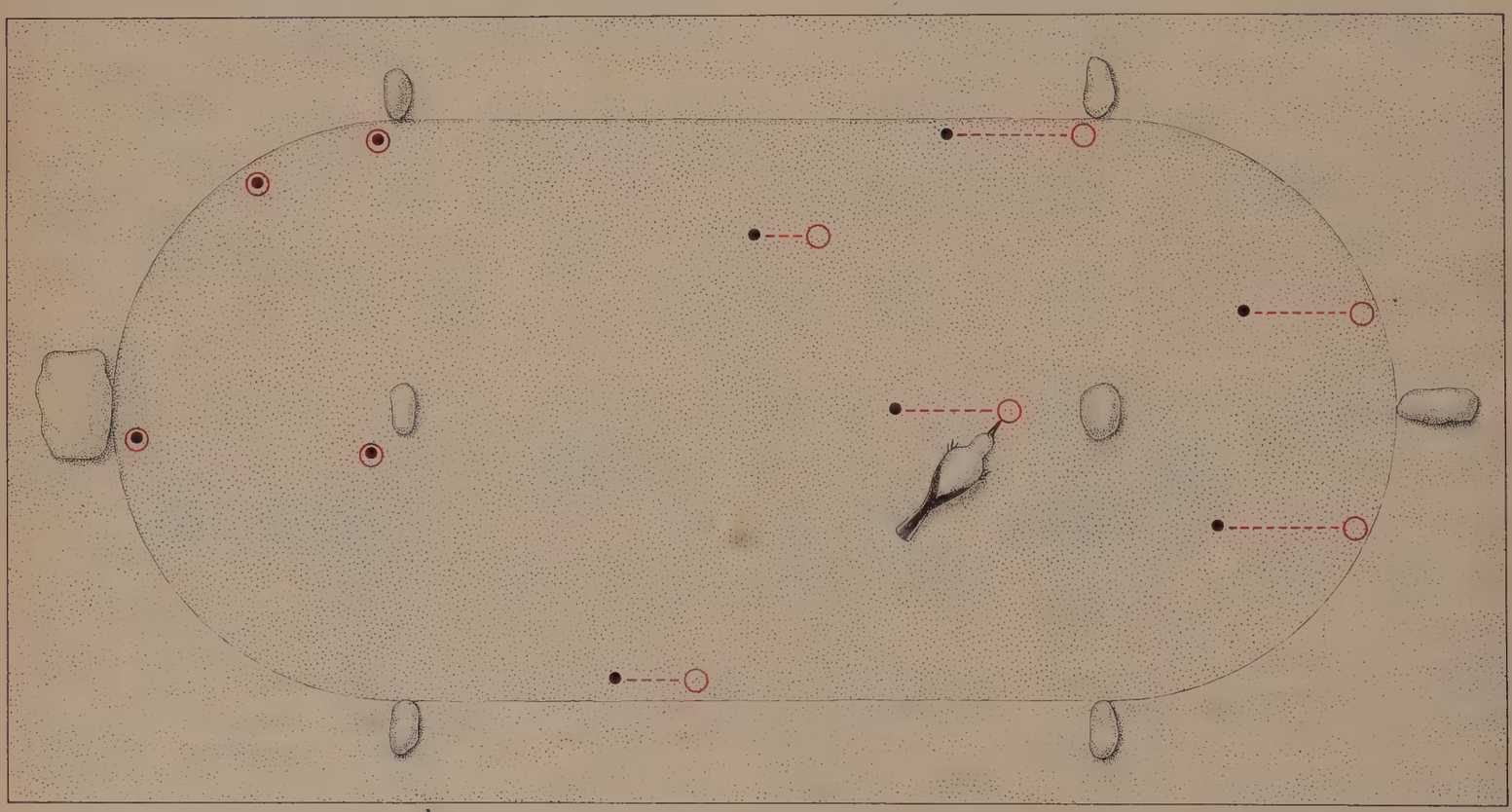


HOARDING OF FOOD by ■ Clark's nutcracker is shown in ■ sequence of drawings based ■ observations made by Stephen B. Vander Wall at Utah State University. The nutcracker jabs its bill into the ground to loosen the soil (*a*); then it inserts ■ conifer seed (*b*), rakes soil and grass over the seed (*c*) and places ■ pebble ■ top of the cache (*d*). The pebble apparently serves only to camouflage the site until weather obliterates all traces of the caching. The procedure takes from 10 to 20 seconds. In nature the nutcracker hoards ■ many ■ 33,000 piñon-pine seeds in thousands of widely distributed caches that average four or five seeds each. The bird does its hoarding from September through November and relies ■ memory to find the caches throughout the following winter, spring and summer.

or some similar cue. On the other hand, the performance of the tits in the first experiment does not necessarily mean that they remember storing each seed in a certain hole out of the 100. To some extent each bird in the laboratory grew to have favorite holes in which it stored seeds particularly often. Each bird was likely to inspect its favorite holes when it was searching for stored seeds. The recovery of seeds from such holes could be due to nothing more than the bird's habit of visiting those places both when it stores seeds and when it recovers them. Indeed, this tendency probably explains why the birds' performance was better than chance when we moved the stored seeds.

A closer look at our data showed that a tit did not simply go to the same holes on each test. Even the holes the tit pre-





PROOF OF MEMORY in Clark's nutcrackers emerged from experiments done by Vander Wall. In the experiment shown here a nutcracker was allowed to cache seeds (*black dots*) within an oval area on the sandy floor of an aviary; four stones were arrayed at each end of the oval (*top drawing*). The oval was then lengthened by 20 centimeters and the stones at the right were moved accordingly (*bottom drawing*). Two days later the bird was allowed to search for its caches. The bird dug with success for the caches at the left (*colored circles*). At the right, however, it made errors of approximately 20 centimeters. Evidently it was relying on its memory of the positions of nearby stones. Errors in the middle of the oval suggest the bird estimated the position of some caches with reference to both sets of landmarks.

ferred most were more likely to be inspected during the recovery of seeds if the tit had actually put a seed there. Nevertheless, the problem was fundamental. In conventional laboratory tests of memory the experimenter supplies the information the animal must try to remember. In our tests the birds themselves produced the information when they stored seeds where they chose. Any tendency they might have to look into the same holes on successive visits to the test room would lead them to perform as if their memory for the storage sites were good.

Our next experiment was therefore designed to make the birds' memory work against their tendency to prefer certain storage sites. Again each bird stored seeds, was removed from the room and then readmitted some two hours later. Instead of requiring the bird to search for the seeds during the second session, however, we now allowed it to store additional seeds. To get it to do so we simply left a bowl full of seeds in the room. We reasoned that if the tits remembered which holes already had seeds, they would avoid those holes when they stored a second batch. Conversely, if they inspected the same holes each time they visited the room, many of the holes they inspected on their second visit would already have a seed in them.

The results were clear. The birds almost never inspected holes containing a seed when they were storing a second batch of seeds. When they were hungry, however, and there were no more seeds to store, they showed their usual efficiency in recovering seeds from holes. We were interested to note that a bird storing a second batch of seeds often chose a hole near one it had employed in storing the first batch. Thus the bird seemed to remember and avoid the individual holes, not just particular branches or parts of the room.

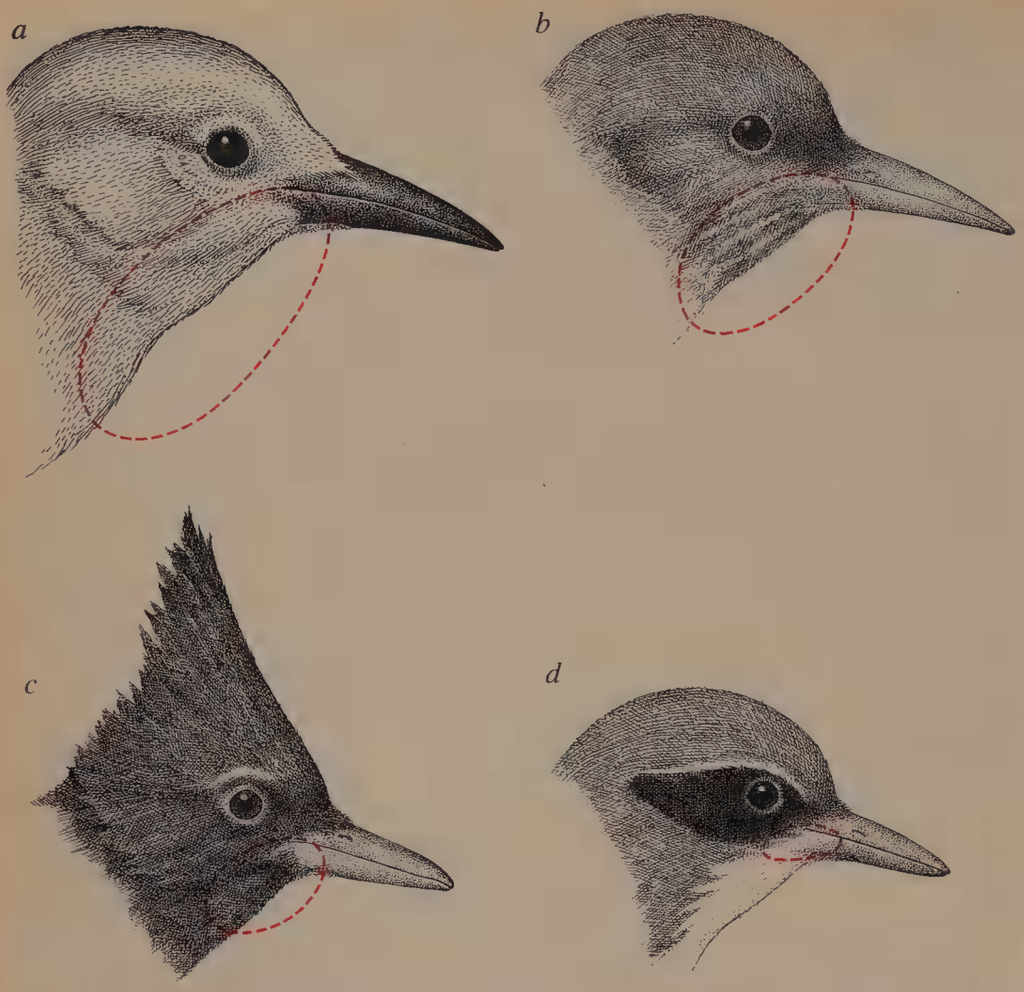
After a bird had stored two batches of seeds, we could let it recover all the seeds. In this way we could test for the recency effect. If marsh tits show the effect, they should recover seeds from the second batch sooner or more often than seeds from the first batch. Our results suggest that there is a recency effect but that it is not strong. If the memory of storage sites fades slowly, however, the two episodes of hoarding would have to be separated by more than two hours for the recency effect to be clear.

An additional finding emerged from the experiments. It is plain that an efficient searcher for stored food should seldom inspect empty sites. Furthermore, it should be able to remember where it has already been to recover food so that it can avoid going back. Laboratory rats faced with the task of collecting food from a number of sites in a large maze prove to have the latter ability. Marsh tits have it too. In the experiments that Krebs and I did, the holes a bird inspected as it recovered seeds were inspected a second time much less than one would expect if the bird were not remembering and avoiding such sites. In experiments done by Sherry, birds tended not to search areas of moss they had already searched for seeds. A marsh tit searching for stored food makes use, then, of two kinds of information: it remembers where food has been stored and also the sites it has already inspected.

SUBLINGUAL POUCH of a Clark's nutcracker is a pocket under the tongue in which the bird carries seeds. It is thus an anatomical specialization that makes the nutcracker well suited for its life of transporting and hoarding seeds. When Vander Wall made this X-ray photograph, the pouch contained 38 seeds with a net weight of 30.6 grams, or more than an ounce.

If information about hoarding sites is stored in the marsh tit's brain, the bird's memory should be subject to a peculiarity of the nervous system that seems to be present in many birds. In many vertebrate animals (including man) information about the left half of the visual field comes from both eyes, but it goes to the right half of the brain, and vice versa. The situation is somewhat different in birds such as the marsh tit, where the eyes are on the sides of the head and each eye surveys a separate visual field. The tracing of information pathways in the brain of the bird and the results of experiments done mostly with pigeons suggest that in these birds information reaching the brain by way of one eye is stored primarily in the half of the brain on the opposite side of the head. In short, such birds have little or no interocular transfer for many kinds of information. Sherry, working with Krebs and Cowie, employed this fact to interpret an intriguing behavior marsh tits show when they store a seed. Having tapped or poked the seed into place, the bird quickly cocks its head first to one side and then to the other. It is as if the bird were looking at the storage site or at the landmarks near the site with each eye in turn.

Does the marsh tit cock its head in order to store visual information about hoarding sites in both halves of the brain? To answer this question Sherry had marsh tits store sunflower seeds wherever they chose in a tray of moss while each bird was wearing a translucent cover over one eye. If the bird was later allowed to search for the seeds with the same eye covered, it found the seeds quite normally. If the patch was moved to the other eye, however, so that the bird had to search with the eye that had not viewed the storage sites, the bird behaved as if it did not remember where the seeds were. If marsh tits do have



ANATOMICAL SPECIALIZATION of four birds in the southwestern U.S. corresponds to their varying degree of dependence on cached food. Clark's nutcracker (*a*) is the most dependent. It is also the most specialized: in addition to a sublingual pouch it has a long, sharp bill, with which it pries open pine cones to get at their seeds. Moreover, it is a strong flier capable of going a great distance burdened with seeds. The piñon jay (*b*) and Steller's jay (*c*) are intermediate in both behavior and anatomy. They carry seeds in their distensible esophagus, not in a specialized sublingual pouch. The scrub jay (*d*) is the least dependent on stored seeds and also the least specialized in its anatomy. It can carry seeds only in its mouth and its bill. In each drawing the seed-carrying capacity of the bird is indicated by a broken colored line.

little or no interocular transfer of information, the behavior is further evidence that they rely on their memory to recover their stores.

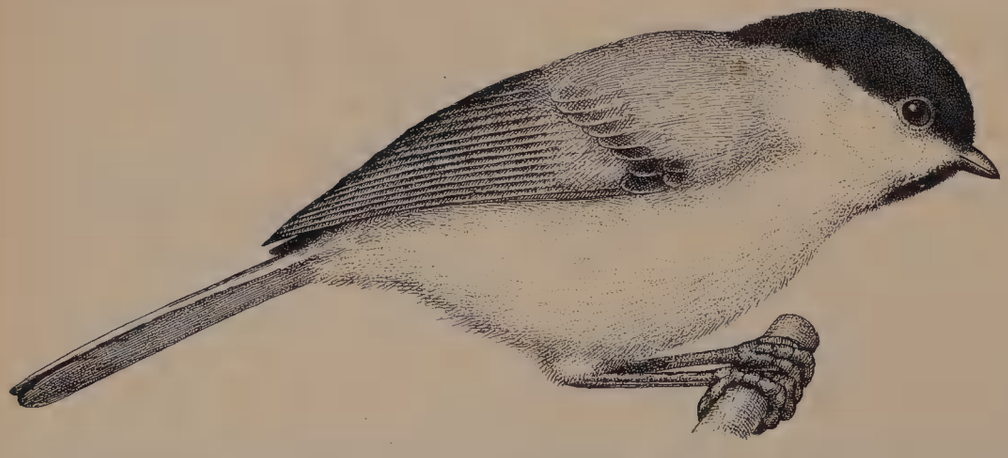
What about other food-storing birds? In the past few years several investigators have been studying birds such as Clark's nutcracker, which makes thousands of winter caches. Again evidence of memory has been sought both in the field and in the laboratory. Of course a field study attempting to discover such things as whether or not a Clark's nutcracker recovers its own caches is extremely difficult to do because of the large number of caches and the long interval between storage and recovery. Few people would want to wait several months near a nutcracker's cache, hoping to learn whether it was eventually dug up by the bird that buried it.

Diana Tomback of the University of Colorado at Denver reasoned, however, that much could be learned from the way Clark's nutcrackers search for caches. She took advantage of the fact that the birds leave a record of their searches in the form of beak marks in the snow and the earth where they dig. Moreover, successful searches can be identified by the presence of piñon-seed coats next to the holes from which seeds were taken. If nutcrackers search at random, successful probe holes and clusters of unsuccessful holes should be more or less evenly distributed over landscapes where the birds have searched. Tomback found, on the contrary, that unsuccessful probes were clumped around successful ones. The pattern indicated the nutcrackers were not searching for caches by trial and error. Furthermore, in the early spring (before rodents had found many of the caches) about two-thirds of the probes were successful, far more than one would expect of a random search.

Tomback's observations do not prove unequivocally that the birds could remember where their caches were. They might have smelled them (although again it is an unlikely possibility). Alternatively, they could have searched mainly in places of the kind that are likely to contain caches. Recent laboratory studies done by Stephen B. Vander Wall at Utah State University seem to rule out these possibilities. Vander Wall studied four captive Clark's nutcrackers in a large outdoor aviary. Two of the birds readily cached seeds in the sandy floor of the enclosure; the other two did not. Nevertheless, the nonhoarders did dig in the sand with their beaks and eat the buried seeds they happened to uncover. Vander Wall allowed the two hoarders to bury seeds in the aviary; then he allowed all four birds to search for seeds.

The results provided dramatic evidence of memory. When the nutcrackers that had hoarded seeds dug in the sand to find them, 70 percent of their probes were successful. When the nonhoarders dug, they found seeds in only 10 percent of their probes. To be sure, a success rate of 10 percent is substantially greater than one would expect for a bird probing completely at random. On the other hand, the two nutcrackers that cached seeds preferred to make their caches near prominent objects in the aviary, such as logs and rocks, and all four birds searched most often near those objects. This tendency accounts for the nonhoarders' success.

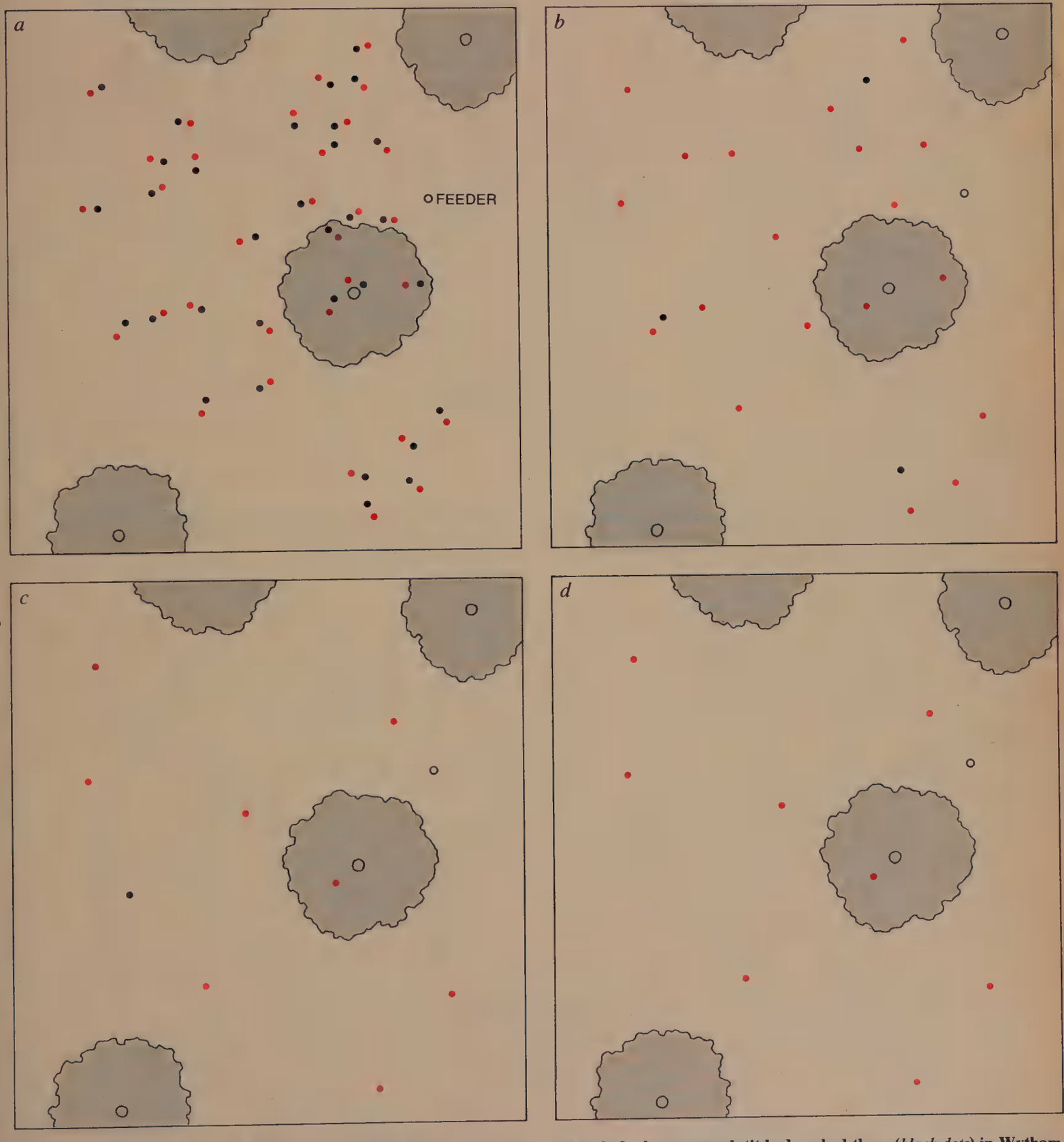
Perhaps nearby prominent objects



MARSH TIT is a small British bird (related to the North American chickadees) that caches seeds if there are more available than it needs immediately for food. Typically the marsh tit returns to draw on its caches within a few hours or at most a few days after it deposits them.

serve as cues for memory when the birds attempt to find their stores. To test this hypothesis Vander Wall covered the aviary floor with a plastic sheet, leaving exposed only an oval area where the sheet had been cut away. At each end of the oval he arranged four large objects such as rocks. The nutcrackers were allowed to make caches wherever they chose in the uncovered oval. Then one end of the oval was extended by 20 centimeters and the four objects at that end were moved in the direction of the lengthening so that they too were each 20 centimeters from their original position.

Would the nutcrackers be able to find caches at the altered end of the oval as well as the ones at the unaltered end, or would their probes be off by 20 centimeters? The latter proved to be the case. Toward the unaltered end of the oval the birds' success rate was high. Toward the altered end the success rate declined. Most of the probes at the altered end were within a few centimeters of the point that was "correct" with respect to the altered position of the nearest large object. The birds also made errors near the middle of the oval, but the errors

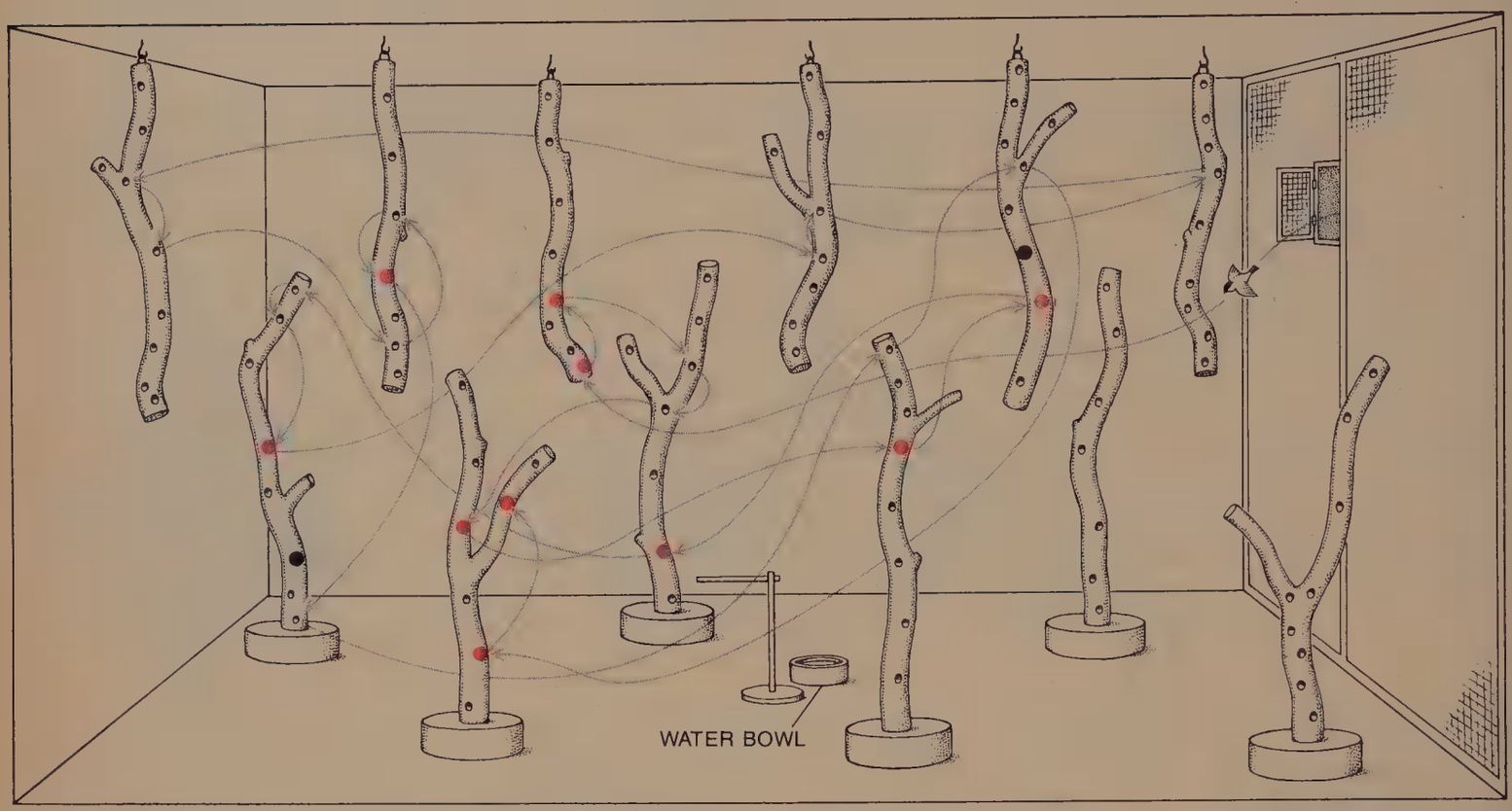


DISAPPEARANCE OF CACHES from Wytham Wood near Oxford was the first evidence confirming that marsh tits use memory to recover hoarded food. Richard J. Cowie, John R. Krebs and David F. Sherry of the University of Oxford coated 30 sunflower seeds with a radioactive substance, then employed a portable scintillation counter to find where a marsh tit had cached them (*black dots*) in Wytham Wood (*a*). One meter from each cache the investigators established a false cache (*colored dots*). Ten hours later (*b*) 19 false caches were undisturbed but only three of the bird's caches remained. The investigators made further checks after 19 hours (*c*) and after 30 hours (*d*).

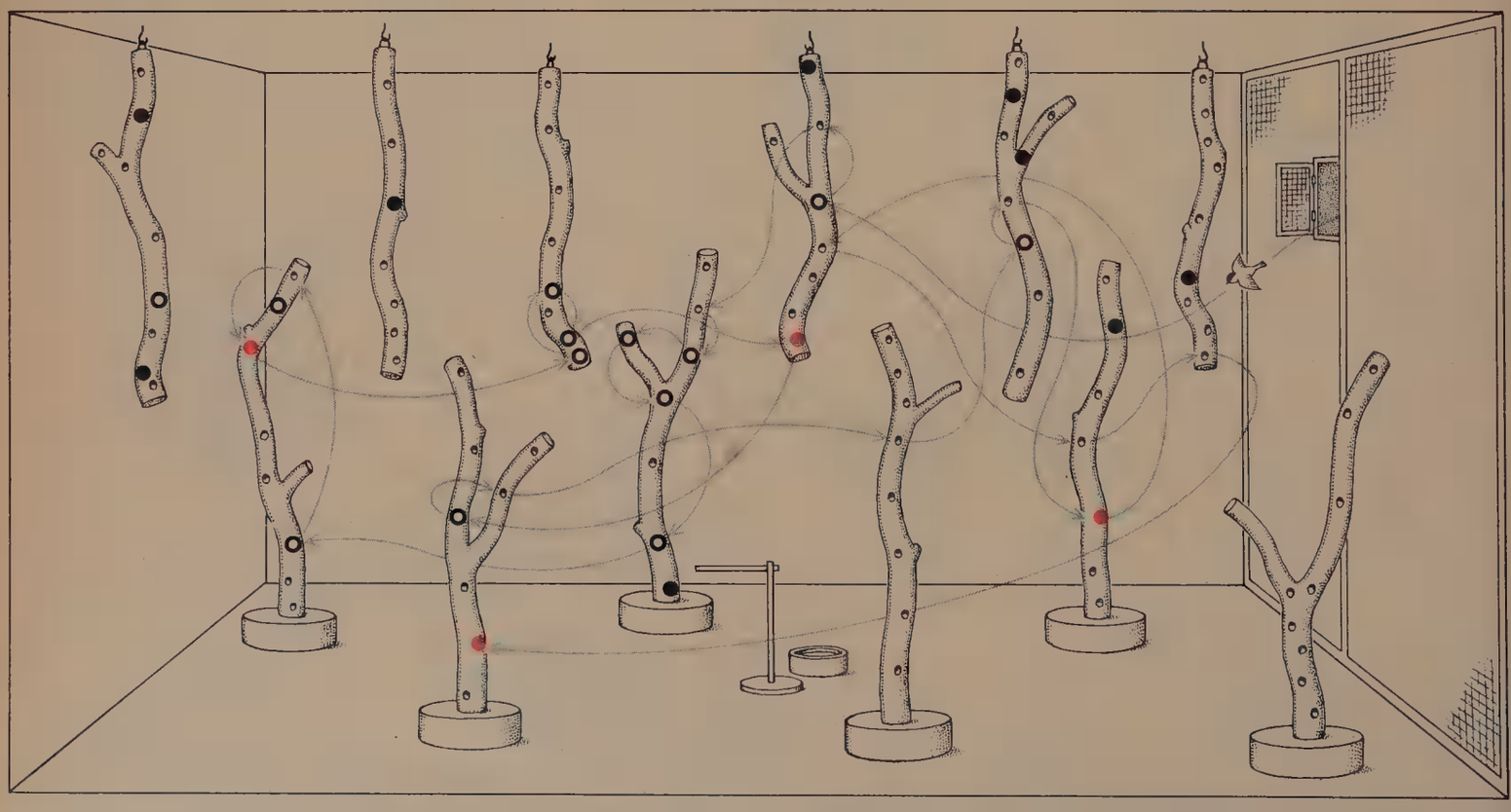
were less extreme. This finding suggests that the cache sites had been established with reference to both sets of objects.

For any species that disperses its stores and relies on memory to recover them one may ask whether the animals' memory enables them to benefit in other ways. That is, are food-storing species better in general at remembering where things are? An attempt to answer this question would require that animals be given tests of spatial memory that do not involve stored food. Such tests have not yet been done with food-storing birds, but thinking about how they might be done reveals some important points.

Among the laboratory techniques employed by psychologists who investigate memory in animals the closest analogue to the hoarding of food is a delayed-



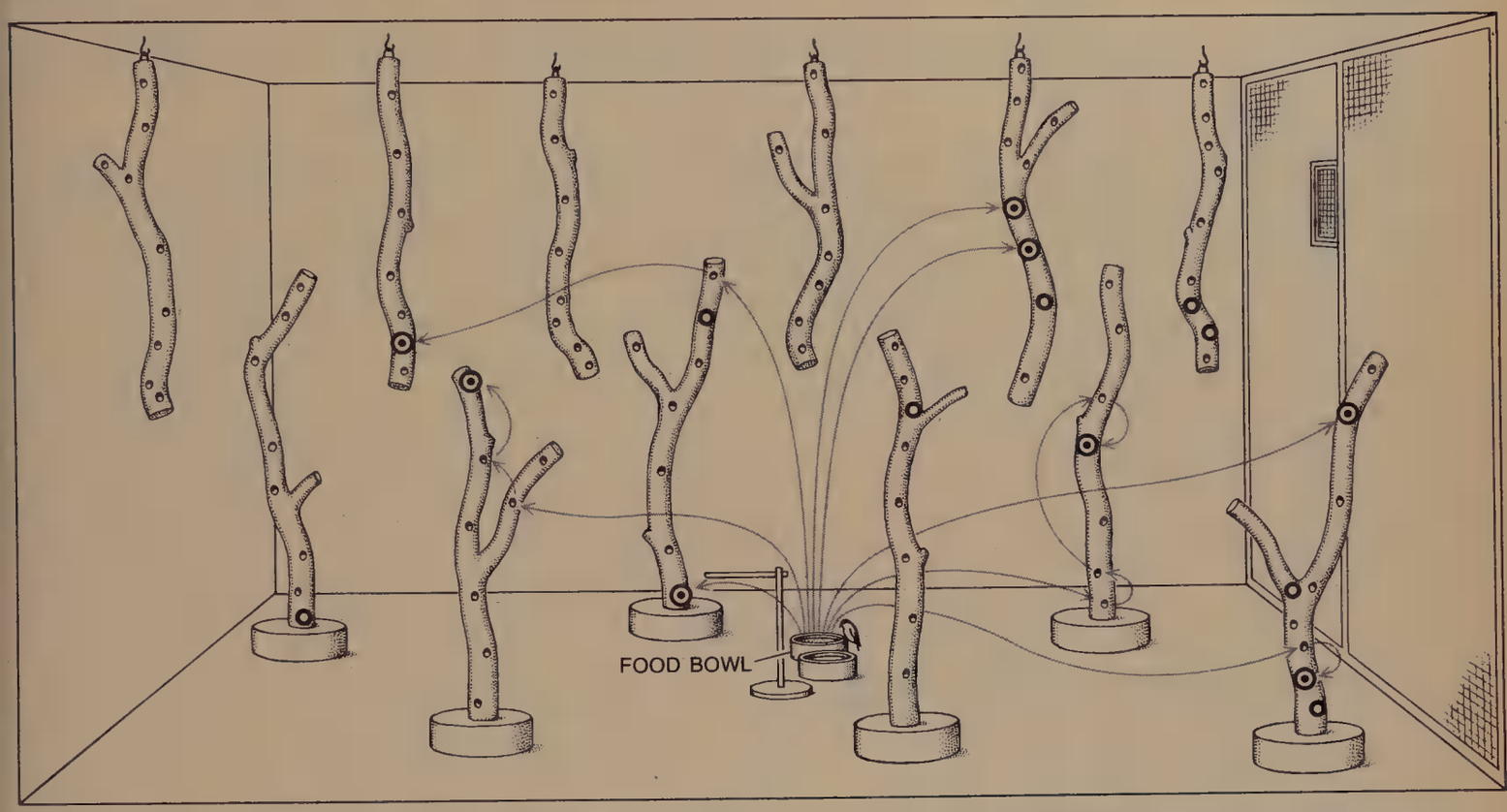
FOUR EXPERIMENTS testing the memory of marsh tits were done by the author in collaboration with Krebs. In each instance a bird stored seeds in some of the roughly 100 holes the experimenters had drilled in tree branches in the laboratory. Here a marsh tit hoarded 12 seeds (*dots*). Readmitted to the room two and a half hours later, it found 10 of the seeds (*colored dots*) by inspecting only 24 holes.



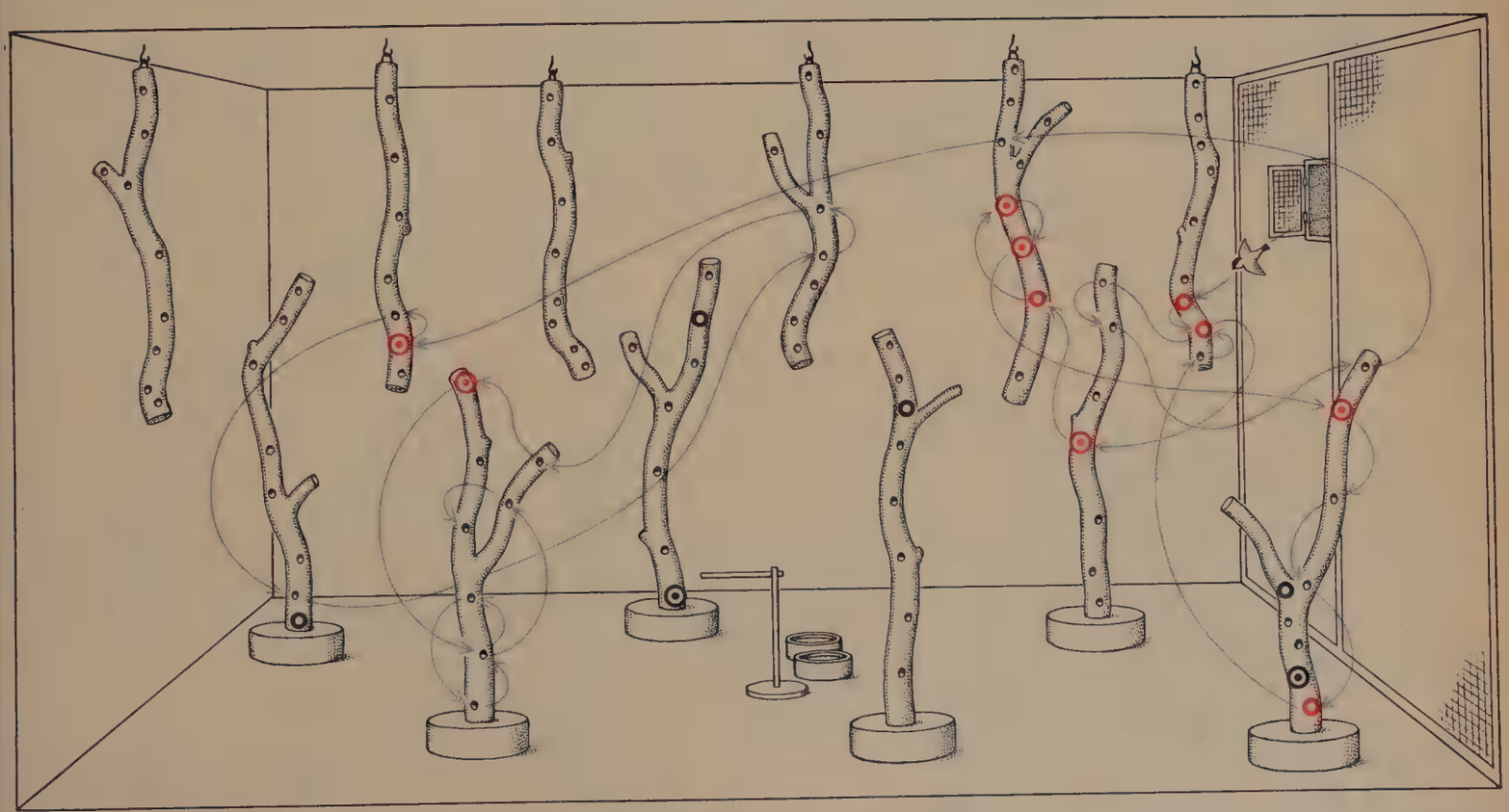
IN A SECOND EXPERIMENT a marsh tit stored 13 seeds (*open black circles*), but Krebs and the author then moved them to other holes (*dots*). Among the first 24 holes the bird inspected when it was returned to the room, 11 were holes it had employed. The experiment indicates that the bird was not detecting the seeds by a cue such as their odor. The bird did succeed in finding four seeds (*colored dots*).

response experiment. Here an animal is shown some visual cue, such as a flashing light or a geometric figure. Later it is offered a choice between the original cue and another one. It is rewarded for choosing the one it saw before. The performance of a laboratory rat or a pigeon typically falls to the level of chance if a few seconds or minutes pass between the first viewing and the second.

Recently Donald M. Wilkie and Russel Summers of the University of British Columbia trained pigeons in a variant of the delayed-response procedure that requires spatial memory. The work illustrates some of the differences between food hoarding and conventional laboratory studies of animal memory. In Wilkie and Summers' experiment a pigeon faced a three-by-three array of nine white disks that could be lighted in-



TWO BATCHES OF SEEDS were hoarded by a marsh tit in the third experiment. In choosing holes for the second batch (*black circles with dots*) it avoided the holes in which it had stored the first batch (*open black circles*). The experiment establishes that the bird was not merely visiting a particular set of holes whether it was storing seeds or recovering them. The marsh tit could remember where it had been.



RECOVERY OF BOTH BATCHES of seeds by the marsh tit that had hoarded them was the final experiment in the series. By inspecting 29 holes it found 10 seeds: four (*open colored circles*) from the first batch of eight and six (*colored circles with dots*) from the second batch of eight. Its first seven inspections of holes were all successes. In two instances the bird made the mistake of revisiting a hole.

dividually from behind. At the start of each trial one disk was lighted briefly. A few seconds later the disk was lighted again, along with another disk or several others. The pigeon was rewarded (it received a pellet of food) if it pecked the disk that had lighted up twice.

The pigeon's task seems to be quite similar to the task facing a food-storing bird. After all, a bird storing food is exposed to visual cues at the storage site, and in order to recover the food it must select the site in preference to similar sites nearby. Yet with only one cue to remember at a time, the pigeons in Wilkie and Summers' experiment appeared not to know which disk had been lighted only a few seconds after they saw it. In contrast, a Clark's nutcracker evidently remembers thousands of storage sites for months.

The pigeon's task did differ in a number of ways from the nutcracker's. The pigeon passively viewed the cue it would be called on to remember. The cues themselves were invariably within a few centimeters of one another and differed only in their position within the array. The food with which the pigeon was rewarded was not behind the cue itself. Any of these factors would be expected to make the pigeon's task more difficult than the task facing a food-hoarding bird. For instance, laboratory investigations done mainly with rats show that animals can remember for several hours the locations they actually visit.

Perhaps, therefore, the types of cues available at hoarding sites are what make the sites memorable. On that hypothesis a food-hoarding bird would not necessarily have a memory extraordinarily better than that of other animals. By adopting a life of hoarding food it would simply have put itself in circumstances where its memory is particularly serviceable. How could this be tested? Comparisons of memory between species that hoard food and species that do not are often compromised by differences in the motivations and the sensory and motor abilities of the species. Obviously the comparisons would be most significant if the species were closely related ones.

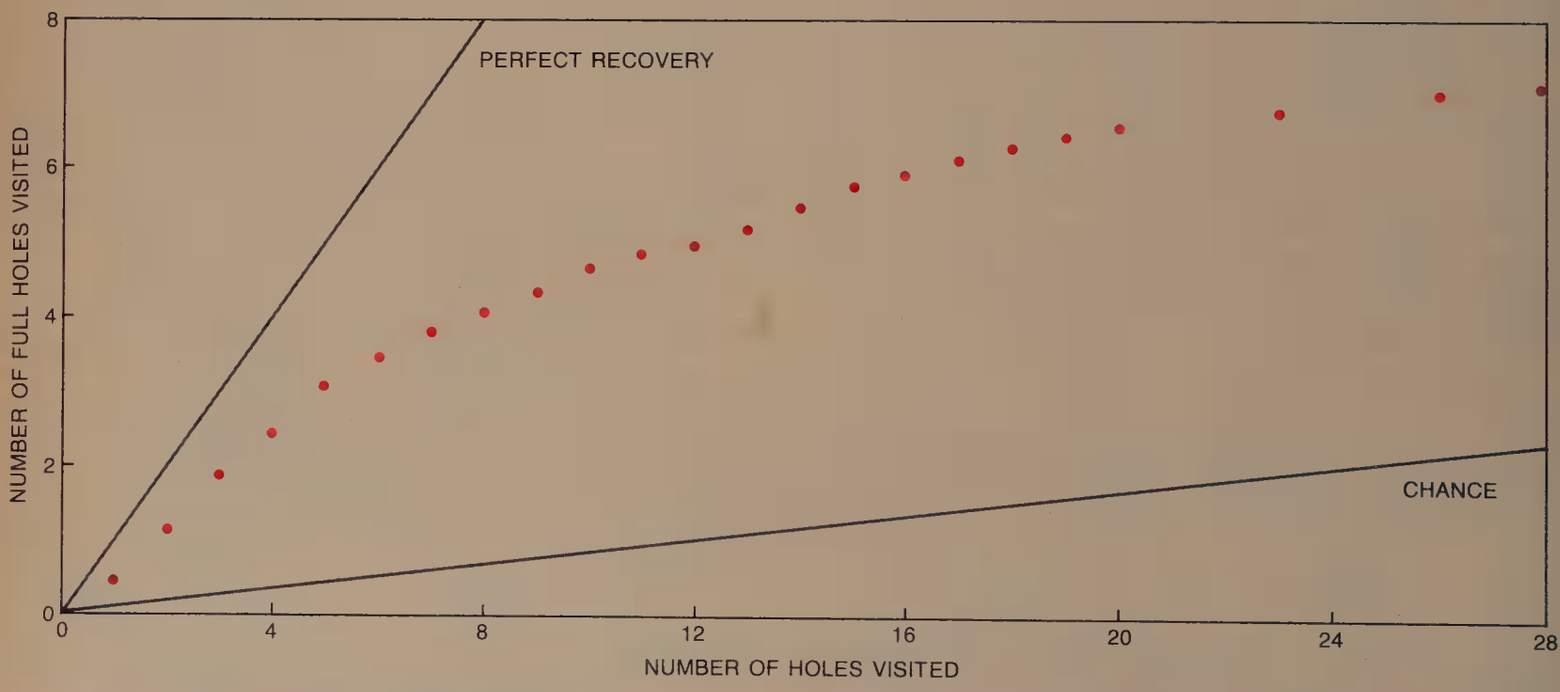
In this respect it is notable that among the corvids (the jays, crows and nutcrackers) and also among the tits some species store food whereas others do not. Among the tits, for example, crested tits and willow tits apparently store food in the fall and draw on it over the winter, much as Clark's nutcrackers do. Marsh tits store superabundant food for relatively short periods. Blue tits and great tits do not store food.

Among the corvids an example of several related species that differ in their tendency to hoard has been described by Vander Wall, working with Russell P. Balda of Northern Arizona University. Clark's nutcracker lives in the same areas of the western U.S. as three other members of the crow family do, namely the piñon jay, Steller's jay and the scrub jay. The four species differ greatly in their dependence on stored piñon seeds, and concomitantly they differ in how well they are able to harvest and transport the seeds.

Clark's nutcracker is the most dependent. By exploiting its stored seeds it is able to begin breeding as early as February, well before most other birds in the area. Clark's nutcracker is also the most specialized anatomically for harvesting and hoarding seeds. In addition to the sublingual pouch in which it carries large numbers of seeds, it has a long, sharp bill, with which it pries open green pine cones before the cones would have opened by themselves. Moreover, Clark's nutcracker is a strong flier. It can carry heavy loads great distances to suitable storage sites.

The scrub jay, in contrast, is the least specialized of the four species. It lacks the anatomical adaptations of a Clark's nutcracker, and it is much less reliant on stored seeds. The seeds it does store it caches entirely within the area where it lives: the area it defends as its own territory. The piñon jay and Steller's jay are intermediate in both anatomy and behavior. Is the spatial memory of the birds of these three species less highly developed than it is in a Clark's nutcracker? The answer awaits further work in the laboratory and the field.

The investigations of memory and food hoarding I have described are at the interface between zoological and psychological studies of animal behavior. Such research is increasing as more psychologists become convinced that it is productive to view laboratory studies of learning and memory as ways of analyzing how animals solve the problems confronting them in nature. At the same time zoologists are becoming interested in aspects of foraging behavior that clearly require a capacity for learning and memory. The joint efforts can therefore combine the zoologist's understanding of how behavior is adapted to the environment with the psychologist's conceptual tools and experimental techniques. The efforts are the beginning of exciting new developments in the study of memory.



MARSH TIT'S PERFORMANCE was charted by Krebs and the author. If the bird were searching at random for eight seeds in 100 holes, it would inspect about 12 holes to find each seed. If its memory were perfect, on the other hand, it would inspect just one hole to find each seed. The bird's performance was intermediate: in 12 trials it inspected an average of about 3.7 holes to find each seed.

THIS CHEVROLET COMES WITH MORE CUBIC FEET AND GOES WITH MORE CUBIC INCHES THAN THE THREE BEST-SELLING IMPORTS.

Our new Cavalier beats the top three import wagons on two important counts: space for people and cargo, and the power to move people and cargo around.

With its new high-compression, electronically fuel-injected 2.0 Liter engine, available new 5-speed transmission and a new lower price,* Cavalier is going to give import wagons some competition they haven't seen before.

Compare Cavalier with any wagon—the quality construction, the load space, the loading ease, the available split-folding back seats, the power, the price, the quickness of its front-wheel drive response. We think it may be just the wagon you're looking for.

See and drive the new electronically fuel-injected Cavalier Wagon, Coupe, Sedan or Type 10 Hatchback today. From America's sales leader. If you haven't seen your Chevy dealer, you're not ready to buy.

Some Cavaliers are equipped with engines produced by other GM divisions, subsidiaries, or affiliated companies worldwide. See your dealer for details.

*Based on a comparison of Manufacturer's Suggested Retail Prices for 1982 and 1983 Cavalier models. Level of equipment may vary.

GM

Buckle up!

Front-wheel-drive Cavalier Wagon.

USA-1 IS TAKING CHARGE

Chevrolet

Oscillating Chemical Reactions

Once thought to violate natural law, reactions where concentrations of key substances periodically rise and fall have now been designed deliberately. They may illuminate similar behavior in living systems

by Irving R. Epstein, Kenneth Kustin, Patrick De Kepper and Miklós Orbán

Oscillating or periodic phenomena are ubiquitous in physics, astronomy and biology. They range from the familiar motion of pendulums and the orbits of planets to the complex biological clocks that govern the daily and seasonal behavior of living organisms. Until fairly recently, however, chemists believed the reactions in their test tubes and beakers were singularly immune to the kind of periodic behavior so familiar in other fields of science. In fact, most chemists trained before 1950 would probably have said that for a mixture of simple inorganic substances to participate in a reaction that oscillated visibly and periodically would be a clear violation of an immutable law of nature. Even today it is usually assumed that chemical reactions are a one-way street: if two substances react to yield a third substance, it is expected that the reaction will continue steadily until the reactants are exhausted or until an equilibrium state is reached. It is not ordinarily expected that the concentrations of the intermediate products of the reaction will reach a certain level, then fall to a lower level, then rise and fall again repeatedly until at some point stable products resistant to further change result.

Although accounts of just such reactions appeared sporadically in the chemical literature of the late 19th century and the early 20th, they were dismissed by the large majority of chemists as being nonreproducible phenomena, attributable perhaps to extraneous processes such as corrosion or film formation occurring in the course of the reaction. In the past 25 years, however, the study of oscillating chemical reactions has finally won respectability and now constitutes one of the most rapidly growing branches of chemistry. The study promises to yield fresh insights into chemical dynamics and the mechanism of catalysis, and possibly into baffling periodic phenomena observed in biology and geology. Whereas the original small group of oscillating chemical reactions was found by accident, it is now possible to describe conditions conducive to oscillation and to mount a systematic search for new oscillating systems. It is a search that has already been rewarded.

The reluctance of chemists to accept the reality of oscillating reactions can be traced chiefly to the second law of thermodynamics. In its best-known formulation, achieved by the 19th-century German physicist Rudolf Clausius, the second law states that the entropy, or randomness, of the universe tends to increase. Applied to chemical reactions the principle requires that a chemical system, in the absence of external inputs of matter or energy, must continuously approach an ultimate equilibrium state. That is, if A goes to B, it must do so throughout the course of the reaction without detours back to A along the way. Reactions where this rule appears to be violated must result, it was thought, either from poorly controlled experimental conditions or even from deliberate deception, since they would constitute a kind of chemical perpetual-motion machine.

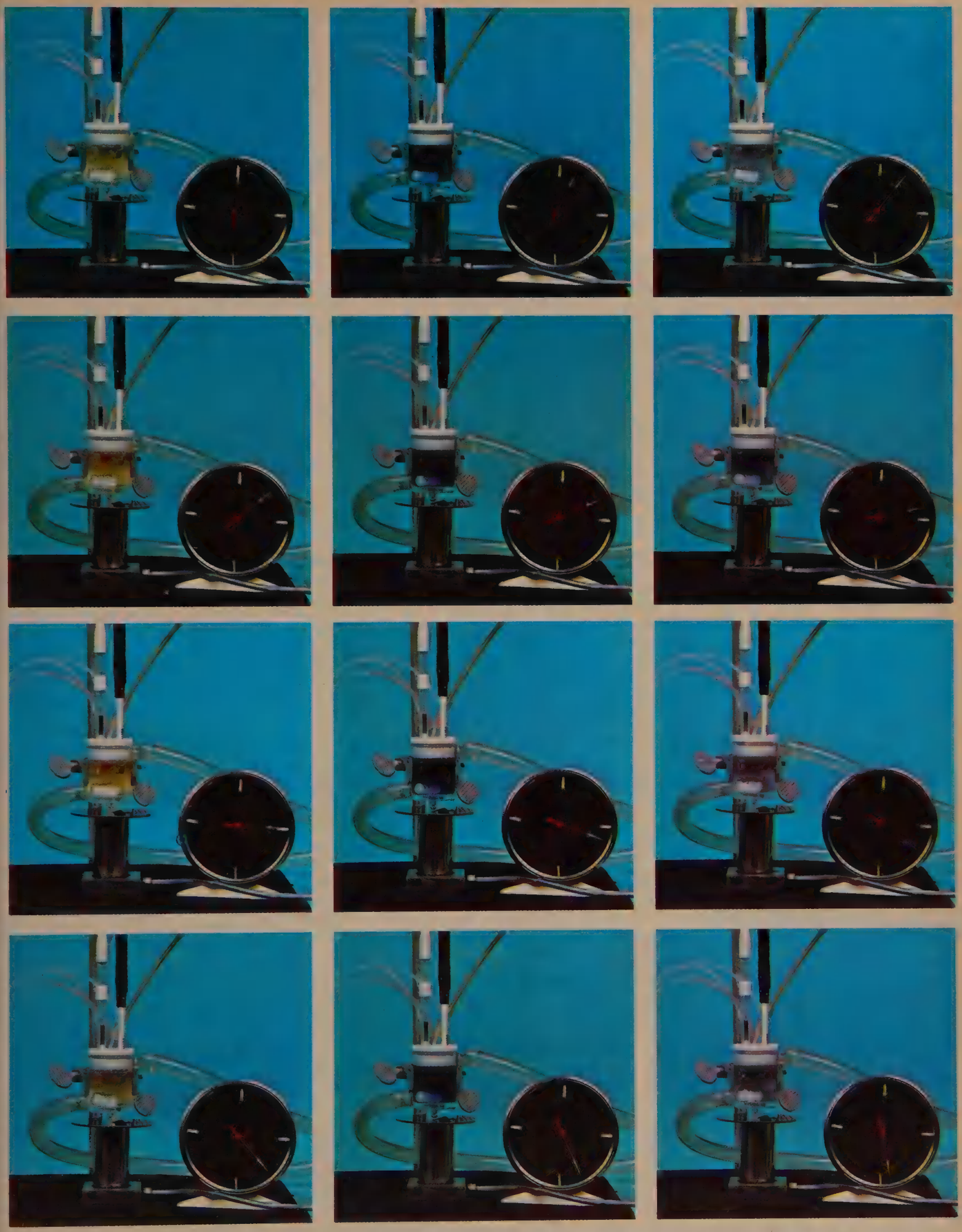
One can therefore imagine the indifference that greeted the report of an oscillating reaction published in 1921 by William C. Bray of the University of California at Berkeley. In studying the role of iodate, the ion of oxidized iodine (IO_3^-), in catalyzing the decomposition of hydrogen peroxide to water and oxygen, Bray noticed that under certain conditions the rate of oxygen production and the concentration of iodine in the solution changed periodically. Bray's observation received scant attention for the next half century. The few scholarly papers that did appear on the Bray reaction were devoted largely to explaining away the oscillations as artifacts due to dust particles or impurities rather than to attempting to examine the phenomenon further.

As we shall see, iodine, the iodide ion (I^-) and the iodate ion play a major role in several of the newly designed oscillating systems. The modern era in the study of oscillating chemical reactions can be dated from an accidental discovery in 1958 by the Russian chemist B. P. Belousov. He noticed that if citric acid and sulfuric acid are dissolved in water with potassium bromate and a cerium salt, the color of the mixture changes periodically from colorless to pale yellow. By 1958 some physical chemists, at least, were prepared to take the report seriously, both because the Belousov reaction was easy to repeat and in part because the 19th-century concepts of thermodynamics had been extended considerably in the years following World War II.

A leader in this effort was Ilya Prigogine of the Free University of Brussels, who recognized that the classical thermodynamics of Clausius required not only that systems be isolated from their environment but also that they be near their equilibrium state. For systems far from equilibrium, either because a reaction is only in its early stages or because the system is "open," hence subject to a flux of energy or matter from the outside, Prigogine and his co-workers developed the concept of an irreversible thermodynamics. For this work Prigogine received the 1977 Nobel prize in chemistry.

In systems that are far from equilibrium a number of new phenomena, described as dissipative structures, can arise. Such structures include periodic oscillations in the concentrations of intermediate species in a chemical reaction; the initial reactants and final products, however, are not subject to oscillation. Living systems are the most interesting and varied examples of open, far-from-equilibrium oscillators. They are maintained in a nonequilibrium state by the input of reactants (nutrients) from the external environment and the output of products (wastes). If either of these fluxes is stopped, the organism and its oscillations must die.

In 1958 the implications of Prigogine's irreversible thermodynamics were understood by only a handful of chemists. Moreover, Belousov's discovery went almost unnoticed because it was



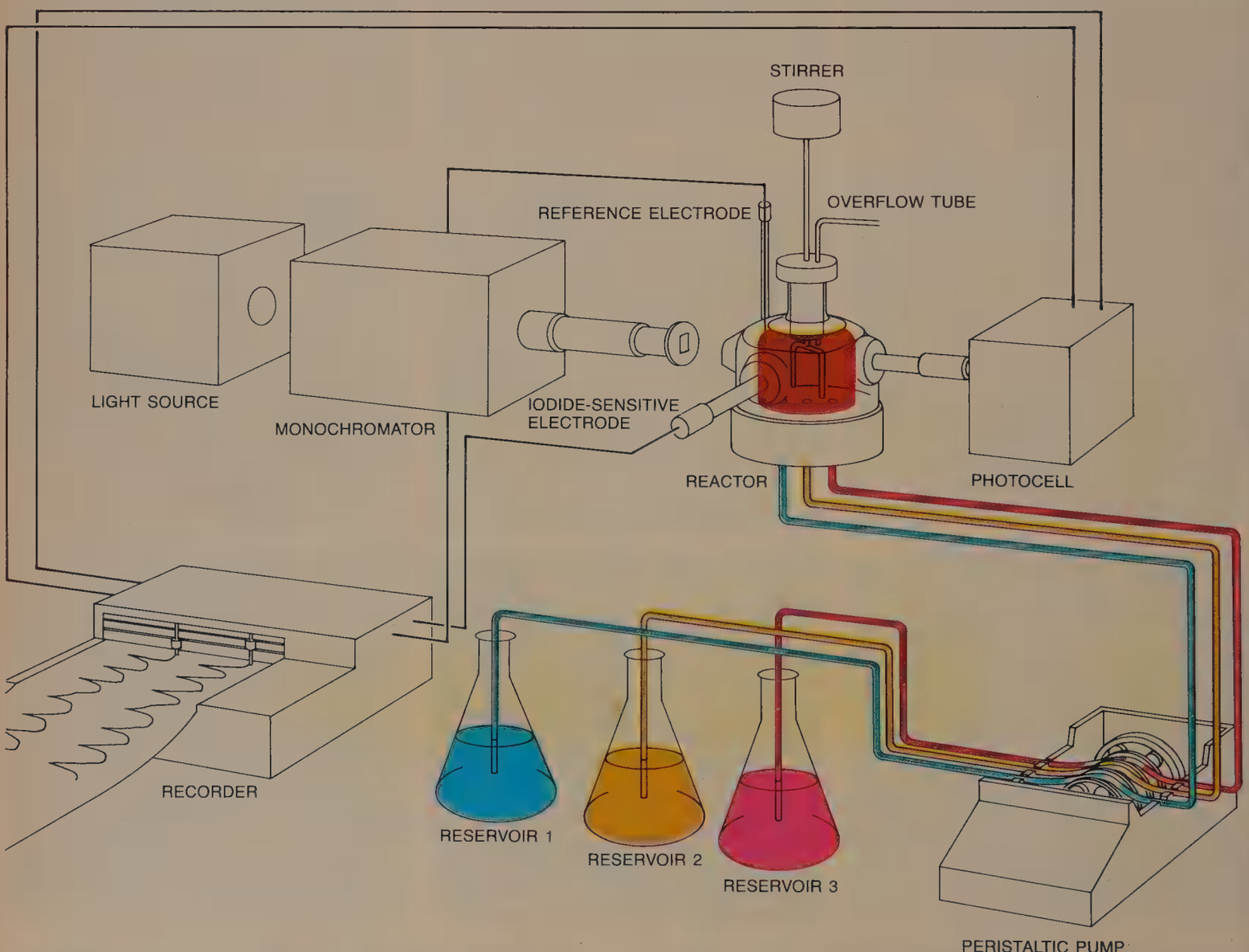
OSCILLATING CHEMICAL SYSTEM is depicted in the authors' laboratory at Brandeis University. The reaction is proceeding in a glass vessel into which three solutions are being pumped at a steady rate: potassium iodate, a solution of perchloric acid and hydrogen peroxide, and a solution of malonic acid containing manganous sulfate monohydrate. Starch serves as an indicator, forming a dark blue complex in the presence of iodine. The interval between successive photographs, beginning at the top left and reading from left to right, is just under three seconds (as is indicated by the second hand on the clock). This particular reaction system was devised by two California high school teachers, Thomas S. Briggs and Warren C. Rauscher. The reaction has been modified to run under continuous-flow conditions.

published in an obscure Russian collection of technical abstracts in radiation medicine. Belousov's oscillating reaction had important advantages over the earlier Bray system. It ran at room temperature and it did not generate noxious products. Furthermore, the oscillations were clearly visible as the cerium changed back and forth from an oxidized (ceric) state, which is yellow, to a less oxidized (cerous) one, which is colorless.

A few years after its publication Belousov's reaction caught the attention of A. M. Zhabotinsky of the Institute of Biological Physics near Moscow. He modified Belousov's recipe somewhat (for example by replacing the cerium reagent with an iron reagent that gave a more distinctive red-blue color change) and began a systematic study of the reaction, which is now commonly called the BZ (for Belousov-Zhabotinsky) reaction. In the early 1960's Zhabotinsky published a large volume of experimental results, including the finding that if a thin layer of initially uniform red BZ solution is allowed to remain undisturbed, blue dots soon appear and grow into a striking array of ring-shaped structures.

Knowledge of the BZ reaction (and of Prigogine's irreversible thermodynamics) spread rapidly in the late 1960's and early 1970's. Although chemists all over the world were fascinated by the BZ phenomenon, the great majority still regarded it as a curiosity, more useful for astonishing students in lectures than as a subject for serious study. In a short time many variants of the BZ reaction were developed by replacing one or more of the components with a closely related chemical species to yield a slower, faster or differently colored oscillator. The most visually impressive of the variants was a hybrid of the BZ and Bray systems devised in 1973 by two San Francisco high school teachers, Thomas S. Briggs and Warren C. Rauscher. Their system, which involves hydrogen peroxide, potassium iodate, perchloric acid, malonic acid, manganese sulfate and starch, changes from colorless to gold to blue and back again. (The detailed recipe is given in the *Scientific American* department "The Amateur Scientist" for July, 1978.)

Although the BZ reaction and its descendants had become well known by the late 1960's through lecture demonstrations, the nature of the oscillatory phenomenon remained a mystery. A chemist understands a reaction only when he can write a mechanism for it. A mechanism is a set of component reactions, called elementary steps, each of which describes an actual encounter between molecules. The familiar "stoichiometric" equation that chemists usually write to describe a reaction shows only the net result, not how the reaction actually proceeds on a molecular scale. For example, when acetylene (C_2H_2)



CONTINUOUS-FLOW STIRRED-TANK REACTOR (CSTR) facilitated the authors' search for oscillating chemical reactions by providing the nonequilibrium conditions necessary for such reactions to take place. The illustration shows a typical experimental arrangement. The reactants, held in three reservoirs, are pumped continuously into the base of the reaction vessel. In the first oscillating system discovered by the authors the feed materials were acidic aqueous solutions of chlorite, iodate and arsenite. The periodic color changes are monitored by the photocell. Oscillations in the concentration of particular ions (here iodide) are sensed by an electrode.

combines with oxygen (O_2) in a welder's torch to form carbon dioxide (CO_2) and water (H_2O), the overall reaction can be written $2C_2H_2 + 5O_2 \rightarrow 4CO_2 + 2H_2O$.

It is wildly improbable, of course, that two acetylene molecules and five oxygen molecules would ever collide simultaneously and fly apart to give four molecules of carbon dioxide and two of water. The reaction actually takes place through a sequence of elementary steps, each one involving a collision between two molecules or the breakdown of a single molecule, often creating or destroying intermediate species that do not appear at all in the equation of the net reaction. Unraveling such a mechanism is an enormous challenge calling for considerable ingenuity, the analysis of much experimental data and in many instances extensive computer simulation.

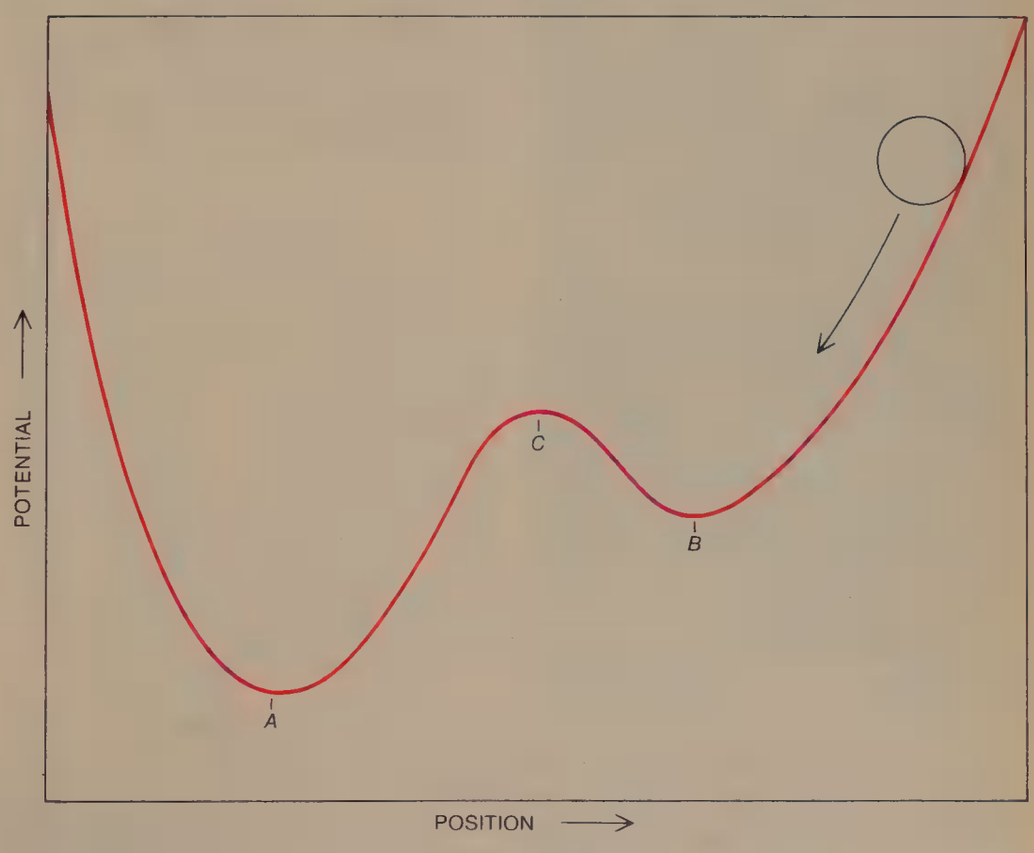
The formidable task of constructing a mechanism for the oscillating BZ reaction was undertaken in the early 1970's by Richard M. Noyes of the University of Oregon in collaboration with Richard J. Field (who is now at the University of Montana) and Endre Körös, a visiting scholar from Eötvös University in Budapest. By 1972 they had conceived a scheme consisting of 18 elementary steps, involving some 20 different chemical species, that appeared capable of explaining the oscillations. Two years later a detailed computer simulation by Noyes, Field and David Edelson of Bell Laboratories confirmed that the mechanism does predict oscillation. Further studies have shown that the mechanism accounts for the development of spatial structures as well. Noyes has likened their approach in unraveling the mechanism to the "method of [Sherlock] Holmes: 'When all other contingencies fail, whatever remains, *however improbable,* must be the truth.'"

With the solution of the BZ mechanism, oscillating chemical reactions became a respectable and exciting area for those interested in reaction mechanisms and chemical dynamics. Various hypotheses were advanced on the conditions required for oscillation in a chemical reaction, and there was optimism in some circles that an understanding of chemical oscillation might throw new light on periodic behavior in living organisms. No one could succeed, however, in defining a set of necessary and sufficient conditions for chemical oscillation. As the decade of the 1970's drew to a close the number of fundamentally different chemical oscillators known was discouragingly small; all had been discovered by accident; only one was well understood, and no one knew how to find any more.

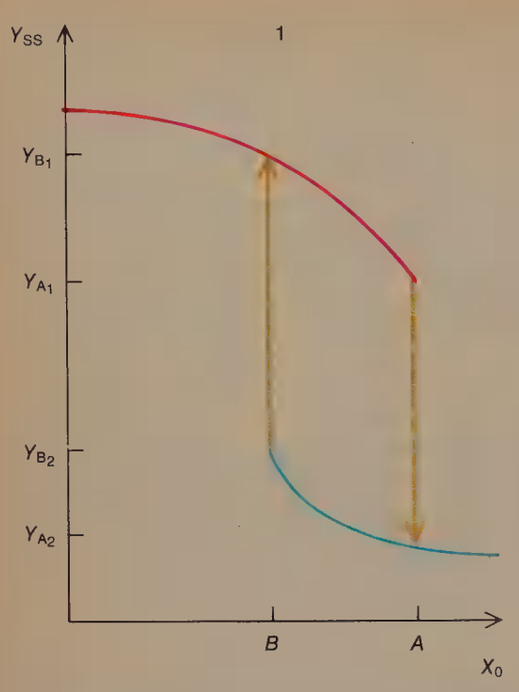
Although biological systems were known to be a fertile source of preparations that exhibit oscillation, they tend

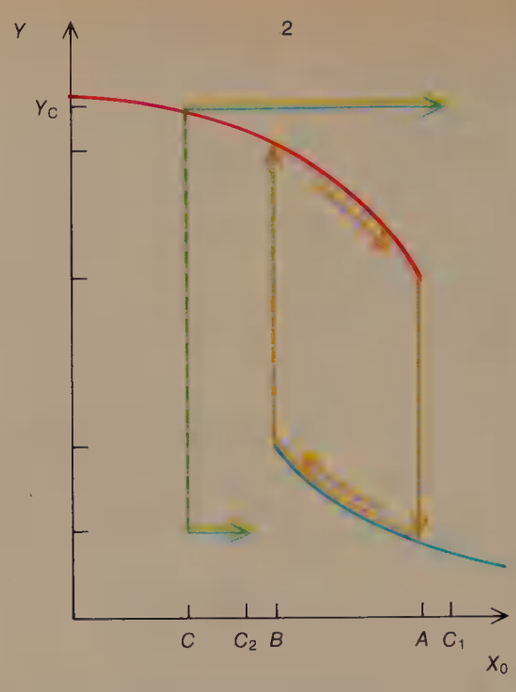
1	$2H^+ + Br^- + BrO_3^- \rightleftharpoons HOBr + HBrO_2$
2	$H^+ + HBrO_2 + Br^- \rightleftharpoons 2HOBr$
3	$HOBr + Br^- + H^+ \rightleftharpoons Br_2 + H_2O$
4	$CH_2(COOH)_2 \rightleftharpoons (OH)_2C{=}CHCOOH$
5	$Br_2 + (OH)_2C{=}CHCOOH \rightleftharpoons H^+ + Br^- + BrCH(COOH)_2$
6	$HBrO_2 + BrO_3^- + H^+ \rightleftharpoons 2BrO_2 + H_2O$
7	$BrO_2 + Ce^{3+} + H^+ \rightleftharpoons Ce^{4+} + HBrO_2$
8	$Ce^{4+} + BrO_2 + H_2O \rightleftharpoons BrO_3^- + 2H^+ + Ce^{3+}$
9	$2HBrO_2 \rightleftharpoons HOBr + BrO_3^- + H^+$
10	$Ce^{4+} + CH_2(COOH)_2 \rightleftharpoons CH(COOH)_2 + Ce^{3+} + H^+$
11	$CH(COOH)_2 + BrCH(COOH)_2 + H_2O \rightleftharpoons Br^- + CH_2(COOH)_2 + HOC(COOH)_2 + H^+$
12	$Ce^{4+} + BrCH(COOH)_2 + H_2O \rightleftharpoons Br^- + HOC(COOH)_2 + Ce^{3+} + 2H^+$
13	$2HOC(COOH)_2 \rightleftharpoons HOCH(COOH)_2 + O{=}CHCOOH + CO_2$
14	$Ce^{4+} + HOCH(COOH)_2 \rightleftharpoons HOC(COOH)_2 + Ce^{3+} + H^+$
15	$Ce^{4+} + O{=}CHCOOH \rightleftharpoons O{=}CCOOH + Ce^{3+} + H^+$
16	$2O{=}CCOOH + H_2O \rightleftharpoons O{=}CHCOOH + HCOOH + CO_2$
17	$Br_2 + HCOOH \rightarrow 2Br^- + CO_2 + 2H^+$
18	$2CH(COOH)_2 + H_2O \rightleftharpoons CH_2(COOH)_2 + HOCH(COOH)_2$

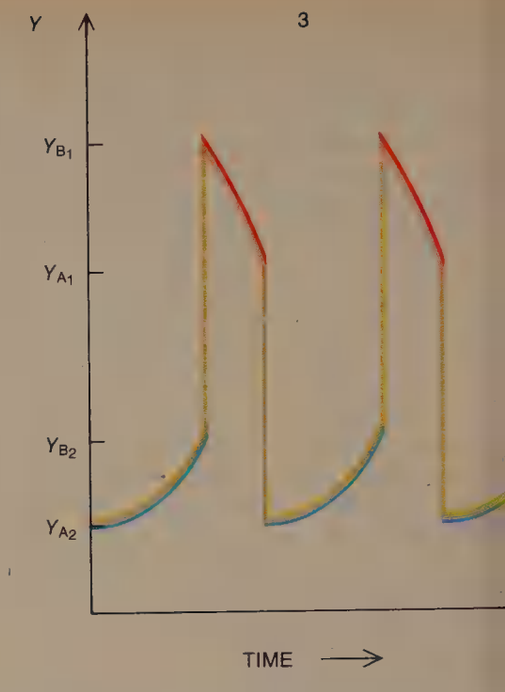
COMPUTER SIMULATION OF OSCILLATION in the Belousov-Zhabotinsky reaction involves 18 elementary steps and 21 different chemical species. The reaction is named for two Russian chemists: B. P. Belousov, who discovered it, and A. M. Zhabotinsky, who improved on it. The starting materials are three inorganic substances, bromate ions (BrO_3^-), bromide ions (Br^-) and cerous ions (Ce^{3+}), and one organic substance, malonic acid ($CH_2(COOH)_2$). A sulfuric acid medium supplies hydrogen ions (H^+). The products of the reaction are carbon dioxide (CO_2), formic acid (HCOOH) and bromomalonic acid ($BrCH(COOH)_2$). As the cerium oscillates between the ceric (Ce^{4+}) and cerous (Ce^{3+}) states the solution is alternately yellow and clear. The reactions shown in color involve only inorganic species and are better understood than the reactions shown in black, which involve species derived from malonic acid. This mechanism was worked out at the University of Oregon by Richard M. Noyes together with his postdoctoral associate Richard J. Field and Endre Körös of Eötvös University in Budapest.



BISTABLE CHEMICAL SYSTEM can be represented by a mechanical analogy. A ball rolling in a potential well with two low points, *A* and *B*, may come to rest in either one. *A* and *B* thus represent stable steady states of the system; *C* is an unstable steady state. Although a ball may come to rest at *C*, the smallest perturbation will topple it into either *A* or *B*. Here the variable is the position of the ball; the external constraint is the shape of the potential well.







HYSTERESIS LOOP is depicted in a bistable chemical system operating in a continuous-flow stirred-tank reactor. As the concentration X_0 of substance X flowing into the reactor is increased from a low value the steady-state value Y_{SS} of species Y in the reactor decreases smoothly along the red curve until at the critical point A it drops suddenly from Y_{A_1} to Y_{A_2}. If X_0 is further increased, Y_{SS} again falls but along the blue curve. If X_0 is decreased, Y_{SS} retraces the blue curve past point A until a second critical point B is reached. Here Y_{SS} jumps from Y_{B_2} to Y_{B_1}. Between A and B either steady state can exist.

OSCILLATION CAN RESULT if a suitable substance Z is added to a bistable chemical system that exhibits hysteresis. If substance Z reacts with substance Y at a low rate to yield substance X, the effective concentration of substance X_0 is modified so that the system "sees" a higher value of X than is actually fed from the reservoir (*left*). The effect of a given amount of Z on X_0 (*green arrows*) will be greater on the red branch of the hysteresis loop (where there is more Y) than on the blue branch. Imagine that the reaction is started with a flow of X equal to C. One now adds enough Z, an amount Z_0, so that the effective X_0 is equal to C_1 when the system is on the red branch and equal to C_2 when it is on the blue. In the absence of Z the flow $X_0 = C$ would yield a steady-state value Y_c. With the addition of Z_0, however, the system follows the red curve as the effective value of X is slowly raised because of the reaction of Y with Z. The system tries to reach the value of C_1 but on reaching the value A it drops abruptly from the red branch and undergoes a rapid transition to the blue branch. There the system "sees" a value of $X_0 = C_2$, and so it drifts to the left along the blue curve in an attempt to match the new effective value of X_0. Again before it reaches that goal, at $X_0 = B$ it suddenly jumps to the red

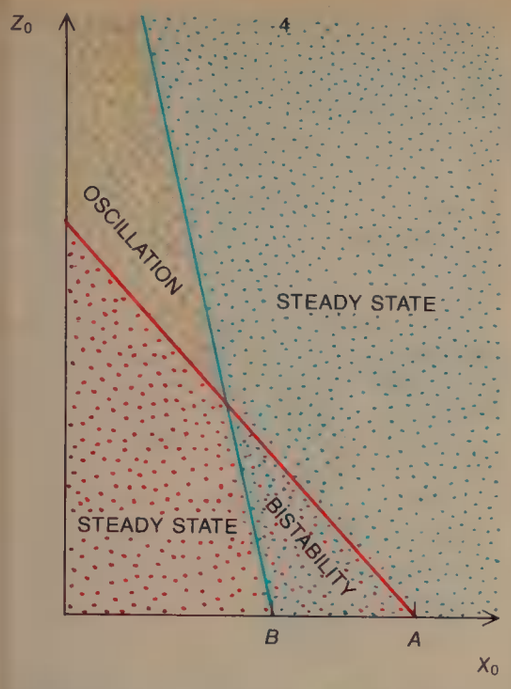
to defy the kind of mechanistic analysis that leads to generalization. Such oscillating systems can be prepared, for example, by extracting the chemical contents of yeast cells and separating them from the cell walls and other extraneous structures. The resulting "soup," when it is fed with appropriate nutrients, exhibits oscillation in the concentrations of hydrogen ions and of the molecule nicotinamide adenine dinucleotide, which in its reduced form (NADH) transports electrons in the cell's normal energy cycle. At the Johnson Research Foundation, Britton Chance and his colleagues have shown that many such biological oscillators involve glycolysis, the process by which cells generate energy by metabolizing sugar in the absence of oxygen. Since such reactions are catalyzed by an array of enzymes, they are daunting in their complexity.

In 1979 two of us (Epstein and Kustin) decided the time had come to try to devise a systematic procedure for designing a chemical oscillator. Where, however, should one start? We began by identifying three conditions that were known to be either necessary for chemical oscillation or conducive to it. The first condition is that chemical systems can oscillate only if they are far from equilibrium. The second is feedback: some product of a step in the reaction sequence must exert an influence on its own rate of formation. The third is that the chemical system must exhibit bistability. Under the same set of external conditions, referred to as constraints, the system must be able to exist in two different stable steady states.

The requirement that the system be maintained far from equilibrium could be met by having the reaction proceed in an apparatus well known to chemical engineers: the continuous-flow stirred-tank reactor (CSTR). In the mid-1970's Adolphe Pacault and his co-workers at the Paul Pascal Research Center in Bordeaux had adapted the CSTR to the study of oscillating chemical reactions. When one of us (De Kepper) left Bordeaux to join our group at Brandeis University in 1980, it was a natural step to build such a reactor. The arrival from Hungary of another of us (Orbán) who had had extensive experience with bromate oscillators completed the group.

A common type of feedback encountered in living systems is autocatalysis, in which the rate at which a substance is produced increases with its concentration. The concept extends to populations of organisms: the number of organisms added to a population in a given interval is almost always proportional to the number of individuals already in the population. (This observation underlies Thomas Malthus' prediction of 1798 that populations tend to outrun food supplies.) In chemistry autocatalytic systems are rare but by no means extraordinary. As early as 1920 Alfred J. Lotka of Johns Hopkins University showed that a simple scheme of two coupled autocatalytic reactions would give rise to oscillation. Although the Lotka mechanism has not proved applicable to any actual chemical reaction, it has guided the thinking of many chemists and has been of considerable utility in describing the oscillations of predator and prey populations in ecological systems.

Let us now consider the role of bistability, or the existence of two different stable steady states, in an oscillating chemical system. By a steady state is meant a condition of the system in which all the variables, such as the concentration of each chemical species, have attained constant values. The steady state is stable if it can adjust to a small change in a variable, say the addition of a drop of acid, and recover without being transformed into a new state. If the slightest change causes a transition to a different state of the system, the original state is said to be unstable. For example, a small ball resting in the bottom of a bowl is in a stable state. A ball



branch and the cycle is completed (*yellow arrows*). X_0 now has the effective value of C_1 again, so that the system moves once more to the right. Repetition of this sequence yields a periodic oscillation in the value of Y (*middle*). The cross-shaped phase diagram (*right*) shows how raising the value of Z_0 can shift the system from bistability to oscillation. If Z equals 0, the system exhibits two steady-state conditions: the "red" state for a small X_0 and the "blue" state for a large X_0. When X_0 lies between A and B, the system is bistable. As Z_0 is increased the region of bistability narrows until at a critical value oscillation begins.

balanced on the rim of the bowl is in an unstable state. In a continuous-flow stirred-tank reactor the external constraints that control the steady-state values in the reaction system are usually the temperature of the thermostatic bath, the concentrations of the chemicals in the reservoirs feeding the reactor and their rates of flow.

If the external constraints on a bistable chemical system are changed, a most peculiar phenomenon known as hysteresis may ensue. The phenomenon is familiar in magnetism, where it takes the form of the "hysteresis loop." When a piece of iron is subjected to increasing magnetizing force, the iron eventually attains full magnetic saturation. If the magnetizing force is then reduced to zero, the iron retains some of its magnetization. In order to reduce the magnetization to zero the iron sample must be subjected to a magnetizing force in the opposite direction. The complete cycle of magnetization and demagnetization takes the form of an S-shaped loop, broad in the middle and pointed at the ends.

An analogous phenomenon can occur in a bistable chemical system. In 1979 a calculation by Jacques Boissonade and one of us (De Kepper) at the Paul Pascal Research Center suggested that oscillations might be induced by adding anoth-

Learn Spanish and French
the U.S. Foreign Service way!

You can be fluent in Latin American Spanish or French. That's right, *fluent!* You can be speaking with all the confidence of a seasoned traveler in a matter of weeks. As well-versed in your new language as if *you* worked for the Foreign Service Corps — because that's exactly whom you'll be learning from.

The Foreign Service Institute has carefully developed, tested, and refined its language courses to train the U.S. State Department's overseas personnel quickly, thoroughly, and effectively. Now the FSI's complete self-instructional audio cassette courses are available to you.

We will send you the entire audio cassette series and instruction books to lead you step by step to fluency in weeks. You work when and where you can, at your own pace, following the taped lessons. You progress according to the schedule *you* establish.

Both courses use native speakers exclusively so that you hear exactly how each word should sound. It won't be long before you start speaking and thinking in your new language. **These programs have to work — our State Department depends on them!**

MONEY-BACK GUARANTEE: Both the Spanish and French courses come with our MONEY-BACK GUARANTEE: try the FSI course of your choice for three weeks. If you're not completely satisfied, return it for a full refund. **There's no risk and no obligation, so order today!**

SPECIAL OFFER: Purchase any language course from Pinnacle and get the **Toshiba KT-S3 Personal Stereo Cassette Player,** complete with headphones, FM Tuner Pack, and shoulder strap (nationally advertised for $139.00), for **only $89.95**

pinnacle Dept. 236, 215 E. 79th St., New York, N.Y. 10021 (212) 772-9202

TO ORDER BY MAIL — Just send this ad to us with your name, address and check or money order. (In N.Y. add sales tax.) Or charge to your VISA, MC, AMEX, or DC account by enclosing your card number, expiration date, and signature.

TO ORDER BY PHONE – (Credit Card Orders Only Please) — Call us Toll Free:
1-800-528-6050 Ext. 1611
In Az 1-800-352-0458 Ext. 1611

Please send me the following course(s):

☐ **Programmatic Spanish — Vol. 1 Basic:** 11 cassettes (16 hrs.) manual and 464-page text. **$100**
☐ **Programmatic Spanish — Vol. 2 Intermediate:** 8 cassettes (12 hrs.) manual and 614-page text. **$90** ☐ **Both Volumes for Only $180.**
☐ **Basic French — Part 1:** 11 cassettes (16 hrs.): and 200-page text. **$100**
☐ **Basic French — Part 2:** 18 cassettes (25 hrs.): and 300-page text. **$125**
☐ **Both Volumes for Only $200.**
☐ Please send the **Toshiba KT-S3 Personal Stereo Cassette Player** at the low price of **$89.95** plus $4.00 for shipping and handling.

SCIENTIFIC AMERICAN

is now available to the blind and physically handicapped on cassette tapes.

All inquiries should be made directly to RECORDED PERIODICALS, Division of Volunteer Services for the Blind, 919 Walnut Street, 8th Floor, Philadelphia, PA 19107.

ONLY the blind or handicapped should apply for this service. There is a nominal charge.

YOUR EXCITING NEW HOBBY!

- Enjoy fantastic savings by assembling your own organ or piano.
- It's easy. No technical knowledge required.
- Just follow our clear, pictured instructions.
- Choose from many models from portables to consoles.
- Ask about our interest-free installment plan.

WERSI
ORGAN & PIANO KITS

WERSI Dept. M41 P.O. Box 5318
Lancaster, PA 17601

☐ Free Info. Pack: ☐ Organ ☐ Piano
☐ Catalog & Demo Record *(enclose $3)*

Name _____
Address _____
City _____ State _____ Zip _____
Phone ( ) _____

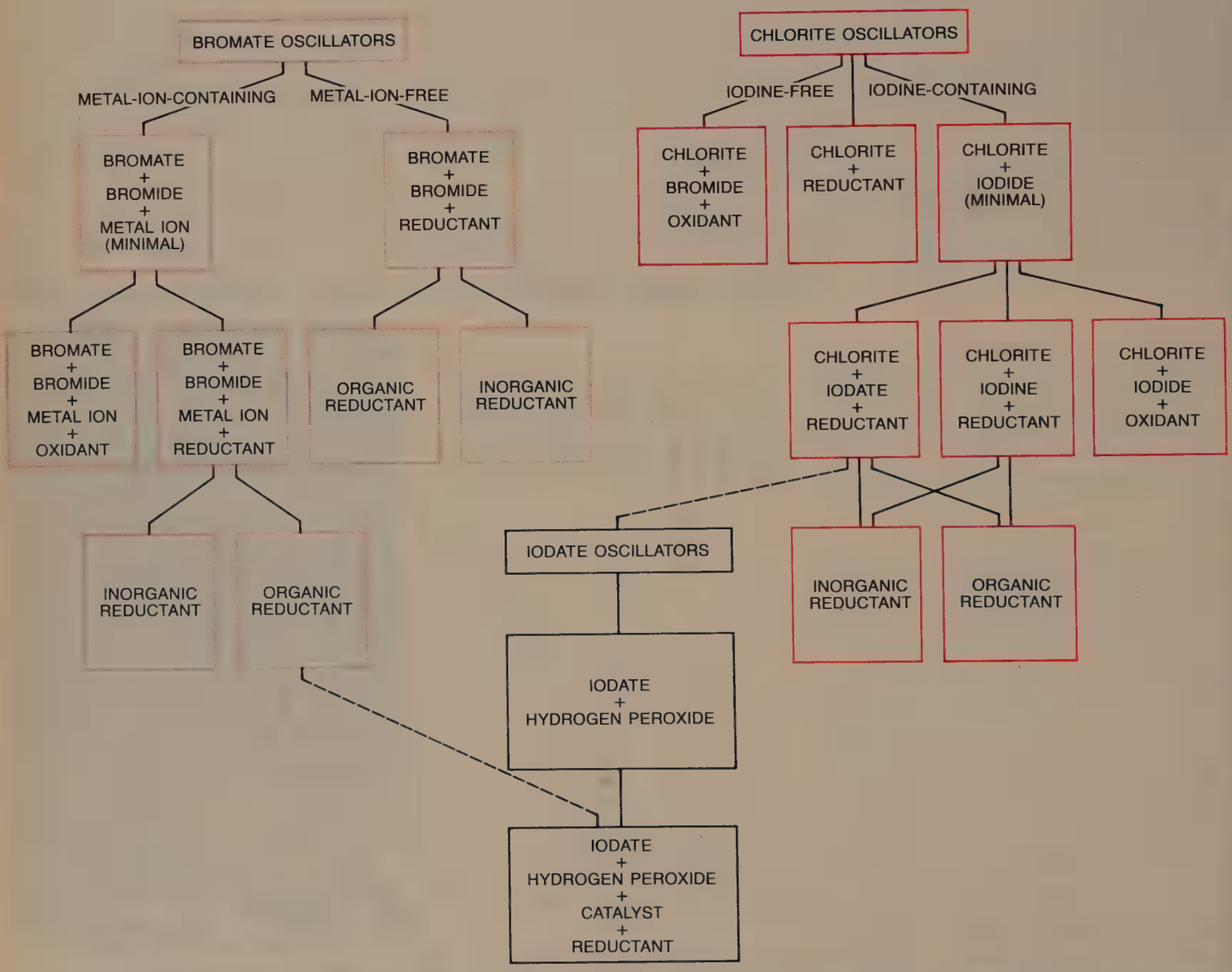
Rep inquiries invited

er chemical to a system capable of hysteresis. The behavior of such a system can be represented in a cross-shaped phase diagram, which shows that if the added substance exerts significantly different effects on the two branches of the system's stable states, oscillations are likely to develop.

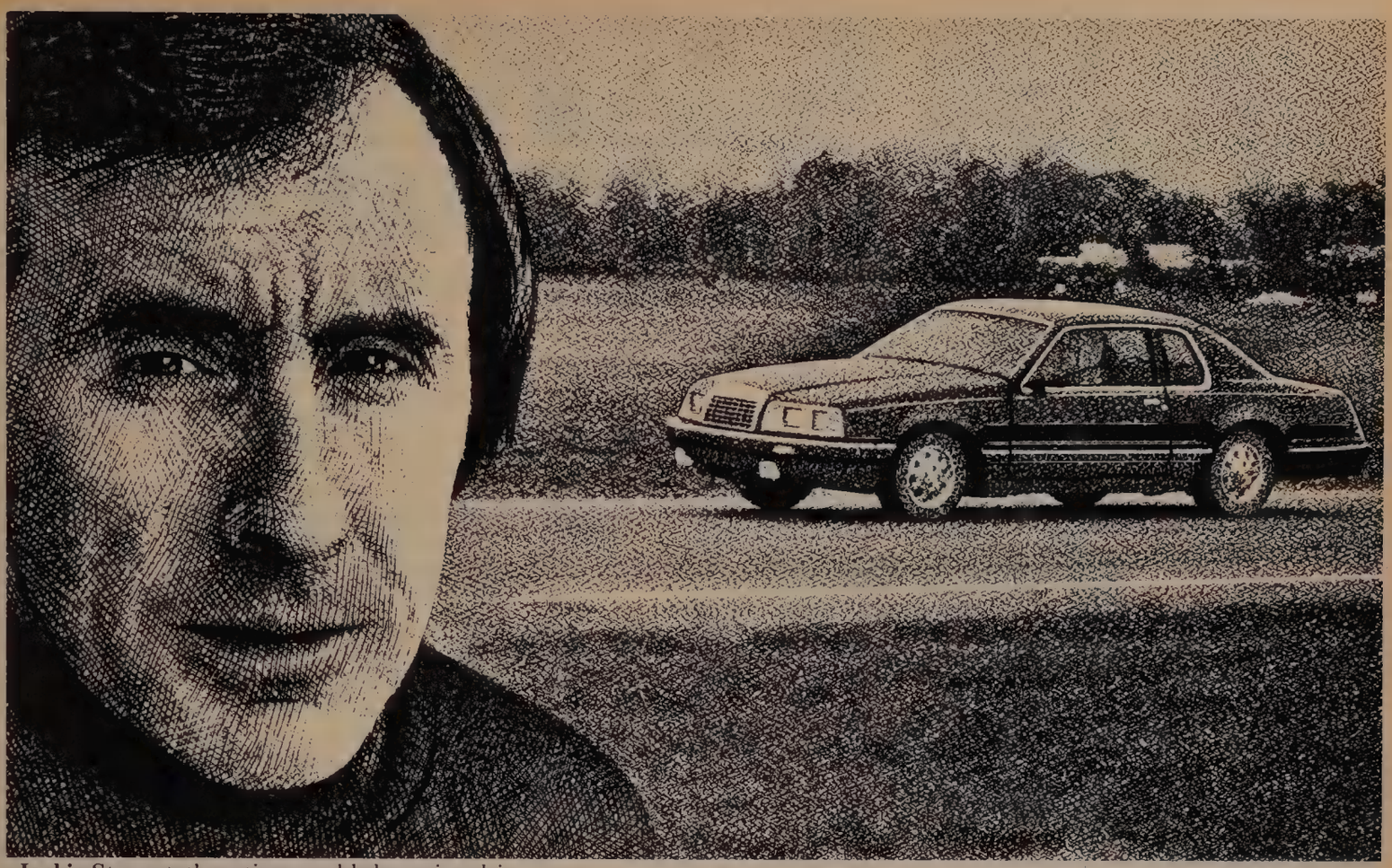
Our plan of attack for designing a chemical oscillator included the following four steps: find an autocatalytic system, run the reaction in a continuous-flow reactor, vary the conditions until a region of bistability is found and then introduce another substance capable of affecting the two branches of bistability differently and thereby of inducing oscillations. Our search of the chemical literature for autocatalytic reactions uncovered several promising ones. Two of them seemed particularly attractive because they shared iodine as a common intermediate. One reaction involved iodate ions and arsenite; a second involved iodide ions and chlorite. Conceivably one reaction might have the desired perturbing influence on the other. The iodate-arsenite reaction soon proved to be bistable. When chlorite was introduced, the composite system began to oscillate almost immediately. The first systematic effort to design a new chemical oscillator was a success.

By varying the components that enter the continuous-flow reactor we have since expanded the initial chlorite-iodate-arsenite oscillator into a family of two dozen related oscillating systems, all having the chlorite ion in common. We can outline a taxonomy of chlorite oscillators, and we are beginning to see how they can be related to oscillators of the bromate and iodate families.

Although bromate oscillators were among the first to be discovered and are far better understood mechanistically than chlorite oscillators, the range of chlorite systems we have found by deliberate design is much broader. Two factors, one chemical and the other historical, account for this seeming paradox. Chemically each family of oscillators has a minimal, or simplest, member from which the others can be derived by adding other substances. The minimal chlorite oscillator, chlorite plus iodide, oscillates over a rather wide range of conditions. In contrast the minimal bromate oscillator, bromate plus bromide plus a metal ion, functions only under the most precisely defined constraints. In fact, even after the existence of the



MAJOR TYPES OF CHEMICAL OSCILLATORS are grouped according to a tentative taxonomy into three main families. The original Belousov-Zhabotinsky oscillator and its many variants are found in the branch of the bromate family that has metal ions. The bromate oscillator in the branch that is free of metal ions is the work of Körös and one of the authors (Orbán) at Eötvös University. All the chlorite oscillators have been discovered by the authors at Brandeis. All the bromate systems involving inorganic reductants were developed at Brandeis under continuous-flow conditions. The iodate family has the fewest members. The first system listed under that heading is the oscillating reaction reported in 1921 by William C. Bray of the University of California at Berkeley and largely ignored for 50 years. The second system in the iodate family is a hybrid of the Belousov-Zhabotinsky and Bray systems. Broken lines connect systems that have certain reactants in common. A reductant is a substance that donates electrons to another substance; an oxidant removes electrons.



Jackie Stewart, three time world champion driver, now consultant to Ford Motor Company.
Get it together—buckle up.

Aerodynamics. It's got to do more than just save gas.

It can if it's properly designed.

For example, aerodynamics can help keep your engine cool, provide a quietness inside the cabin of the car and, more importantly, help make your car handle better.

The tunnel can't lie

There are pretty cars, and aerodynamic cars. Cars that combine both qualities come out of the wind tunnel. There are some visual hints that will help you decide which cars combine both good looks and good aerodynamics.

A few hints

Are the side view mirrors tucked away out of the slip stream? Does the air dam extend well below the front bumper? Are the windshield, hood and front fenders raked back at a steep angle?

When the engineers get it right, you have a car that slips down the highway quietly going about its business without taking too much out of the driver or his pocketbook.

Case in point: the 1983 Ford Thunderbird. It has all those good aerodynamic qualities and, best of all, a drag coefficient of 0.35. One of the lowest of any car in its class.

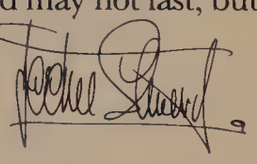
Pretty, but practical

But what I find really interesting is that the people at Ford have done this with a system of aerodynamics that also produced cars that are road worthy and display good manners at highway speeds.

That's because the designers and engineers didn't work in a vacuum. Every one of today's new generation of Ford Motor Company cars had to prove itself in the wind tunnel and on the test track. The results show that it was worth that extra effort.

Light at the end of the tunnel

Words written on the wind may not last, but I have a feeling that cars designed in the wind will be with us for a long, long time.



Ford has it. Now.

Ford

FORD · MERCURY · LINCOLN · FORD TRUCKS

minimal bromate oscillator had been predicted by one investigator, Kedma Bar-Eli of Tel-Aviv University, the search for it was fruitless until two of us (Orbán and Epstein) succeeded in finding it by following our systematic approach based on the cross-shaped phase diagram.

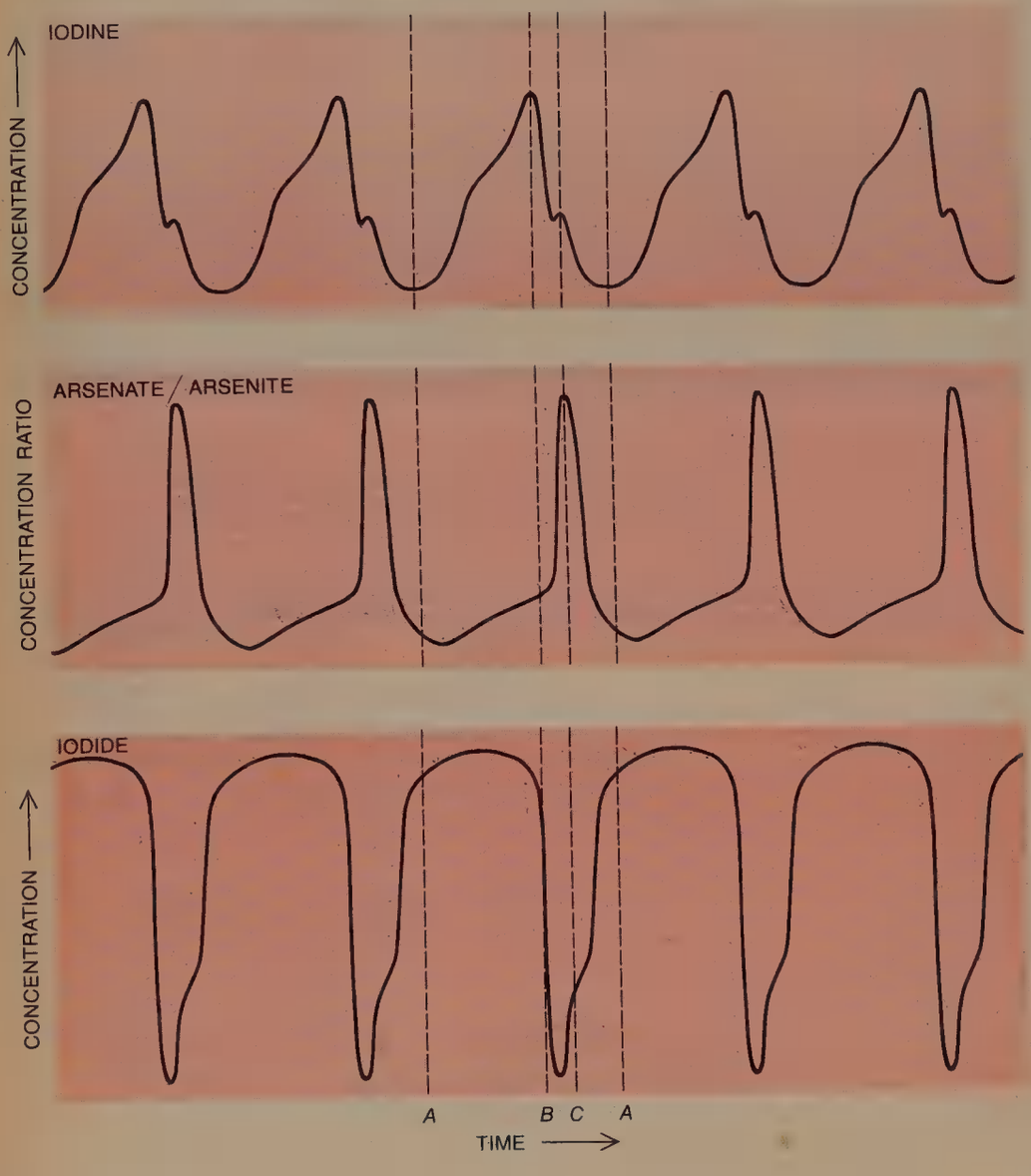
From the historical point of view the first bromate oscillators were discovered before continuous-flow reactors came into service. The bromate systems oscillated readily under batch (zero flow) conditions. Chlorite oscillators, like the vast majority of new systems, show periodic behavior only under continuous-flow conditions because the flow is needed to maintain them far from equilibrium. A system that oscillates under batch conditions will nearly always oscillate under continuous flow, but the converse is rarely true. Most bromate oscillators were originally sought under batch conditions; chlorite oscillators were originally sought only under flow conditions. Accordingly the variety of chlorite systems discovered soon exceeded the variety of bromate systems. (We subsequently found, however, several chlorite oscillators that work under batch conditions.) One might say that the very ease of oscillation of the first bromate systems hindered the search for more of them. It is only quite recently that continuous-flow reactors have been intensively applied to the search for new bromate oscillators and to the study of the third (and so far the smallest) family of oscillators: the iodate systems.

One of the aims of designing a new chemical oscillator was to have a system whose mechanism could be deciphered and compared with that of previously examined oscillating reactions. The dissection of the family of oscillating chlorite systems is now well under way. We have completed studies of the rates of several component processes and thus have several pieces of the puzzle; a complete description of the oscillating scheme now appears to be a matter of turning groups of pieces around until they lock together in just the right way.

We can already give a schematic view of how our first oscillating system functions, the one whose inputs are solutions of either potassium or sodium salts in the form of a chlorite, an iodate and an arsenite. Only four principal overall reactions are involved, each consisting of several elementary steps. Two of them involve the reaction of chlorite with either iodide or iodine, yielding chloride and iodine in the first case and chloride and iodate in the second. In the third reaction iodate and iodide react to form iodine. In the fourth iodine and arsenite form iodide and arsenate. Variations in the rates of these four reactions give rise to periodic oscillations in the concentrations of iodine and iodate and in the ratio of arsenate to arsenite [*see illustration at left*].

Chemical oscillation, although fascinating in itself, is related to several other phenomena of at least equal interest. One is the formation of spatial structures in an initially homogeneous medium. Such structures develop, for example, in one of the oscillators we have discovered where the ingredients are chlorite, iodide and malonic acid. When starch is present as an indicator, the solution is initially a uniform purple, because of a complex formed by iodide, iodine and starch. As the reaction proceeds white dots appear and grow into rings and sets of concentric rings, which annihilate one another when they collide. One observer compared the emergence of white dots against the purple background to seeing stars come out. The appearance of such geometric forms in a previously undifferentiated mass is highly suggestive of the kind of process that may enable embryonic cells in animal organisms to sort themselves out into individual types destined to become blood, brain or bone.

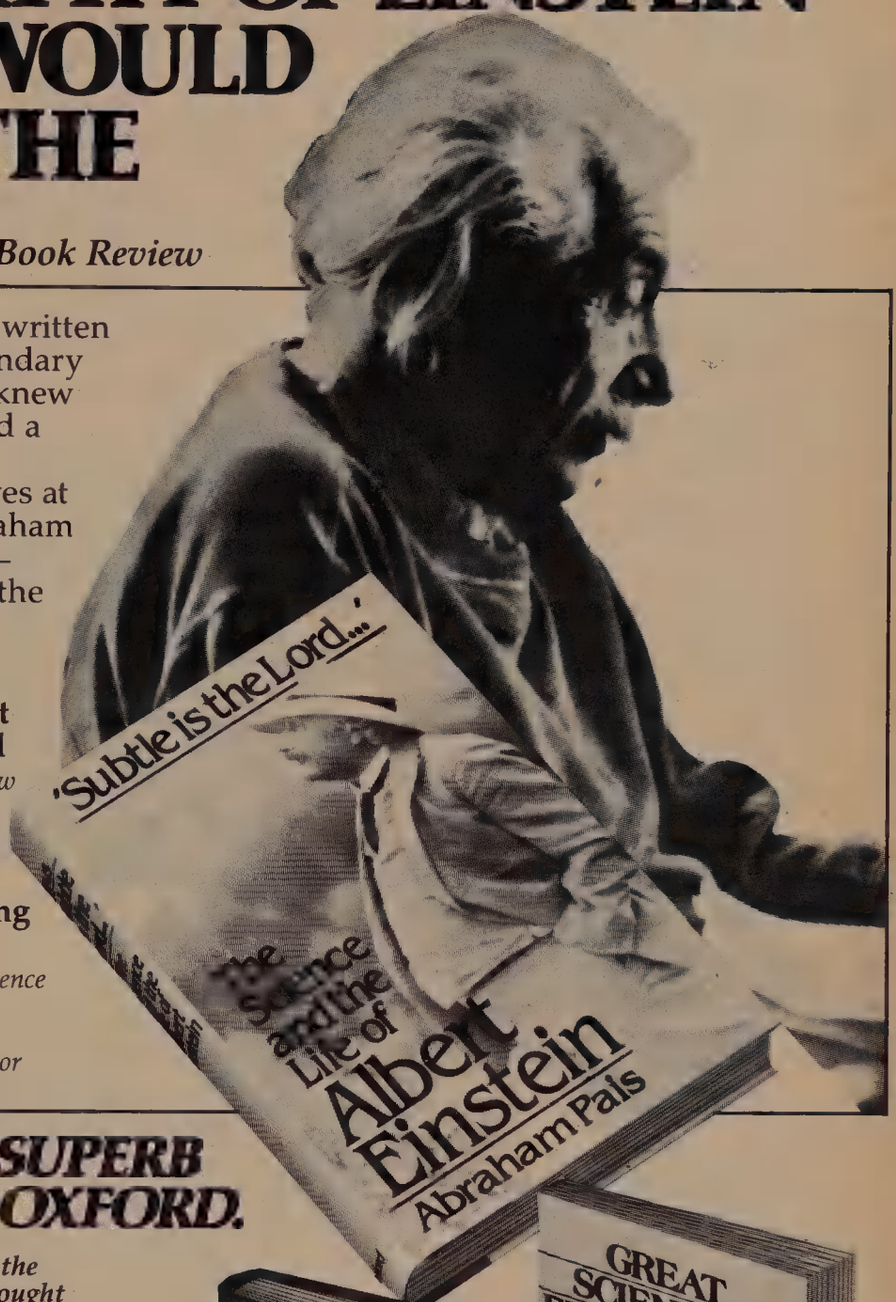
Another related phenomenon, whose



1 CHLORITE + IODIDE → CHLORIDE + IODINE
2 CHLORITE + IODINE → CHLORIDE + IODATE
3 IODATE + IODIDE → IODINE
4 IODINE + ARSENITE → IODIDE + ARSENATE

PATTERN OF OSCILLATIONS in the first oscillating system devised by the authors can be understood qualitatively from the interplay of four main reactions. The oscillating system is supplied with a uniform and balanced flow of three salt solutions: a chlorite, an arsenite and an iodate. Within the continuous-flow reactor the concentrations of iodine and iodate and the ratio of arsenate to arsenite oscillate as is shown by the three curves. At point *A* iodide and arsenite are high whereas iodine and arsenate are low, so that the predominant reactions are *1* and *3*. These reactions sharply increase the concentration of iodine and begin to decrease the concentration of iodide, bringing the system to point *B* and shutting down reactions *1* and *3*. Reactions *2* and *4* now come into play, depleting the iodine and converting arsenite into arsenate. The system reaches point *C*, where arsenite and iodide are at or near their minimum levels. Arsenite is replenished by the input to the reactor and arsenate is depleted by the output. Reaction *4* regenerates iodide while consuming the iodine, thereby returning the system to *A*.

"THE BIOGRAPHY OF EINSTEIN HE HIMSELF WOULD HAVE LIKED THE BEST." —*The New York Times Book Review*



Only an eminent scientist could have written this first scientific biography of the legendary physicist. And only an individual who knew Einstein personally could have produced a portrait of such intimacy and insight.

Aided by access to the Einstein archives at Princeton, award-winning scientist Abraham Pais sets Einstein's revolutionary ideas—his successes as well as his failures—in the context of twentieth-century science.

"Thoroughgoing, accurate, witty and clear as brook water, it is a work against which future scientific biographies will be measured." —*The New York Times Book Review*

"Sympathetic but clear-eyed.... a fine book." —*Scientific American*

"A coherent account of almost everything of scientific significance that Einstein did.... Unique and indispensable." —*Science*

"An extraordinary biography of an extraordinary man." —*Christian Science Monitor*

853907-X, $25.00

ENJOY THESE OTHER SUPERB SCIENCE BOOKS FROM OXFORD.

PLUTO'S REPUBLIC **Incorporating *The Art of the Soluble* and *Induction and Intuition in Scientific Thought***
by Peter Medawar

This definitive collection of Nobel Prizewinner Peter Medawar's work explores the nature of contemporary science—and scientists. "In these erudite and outspoken essays, Medawar gives us the pleasure of watching a lively and nimble mind at play." —*Newsweek*

217726-5, $25.00

MATHEMATICAL SNAPSHOTS **Third American Edition, Revised and Enlarged** by H. Steinhaus, with a new preface by Morris Kline

Using striking photographs and diagrams, Steinhaus unravels hundreds of tantalizing mathematical phenomena—from simple puzzles to challenging equations. "An amazing display of the richness, the variety and especially the interrelatedness of mathematical thought." —*Scientific American*

503267-5, $7.95 paperback

NOW IN PAPER-BACK!

GREAT SCIENTIFIC EXPERIMENTS
***Twenty Experiments that Changed Our View of the World*. First American Edition**
by Rom Harré

From Aristotle's embryos to Galileo's falling bodies to Newton's color spectra—here are 20 brilliant investigations that unlocked nature's secrets. Each breakthrough is vividly retold using original sources, rare illustrations, and the scientists' own words. "Beautifully lucid.... The book is a great success." —*New Scientist*

286036-4, $8.95 paperback

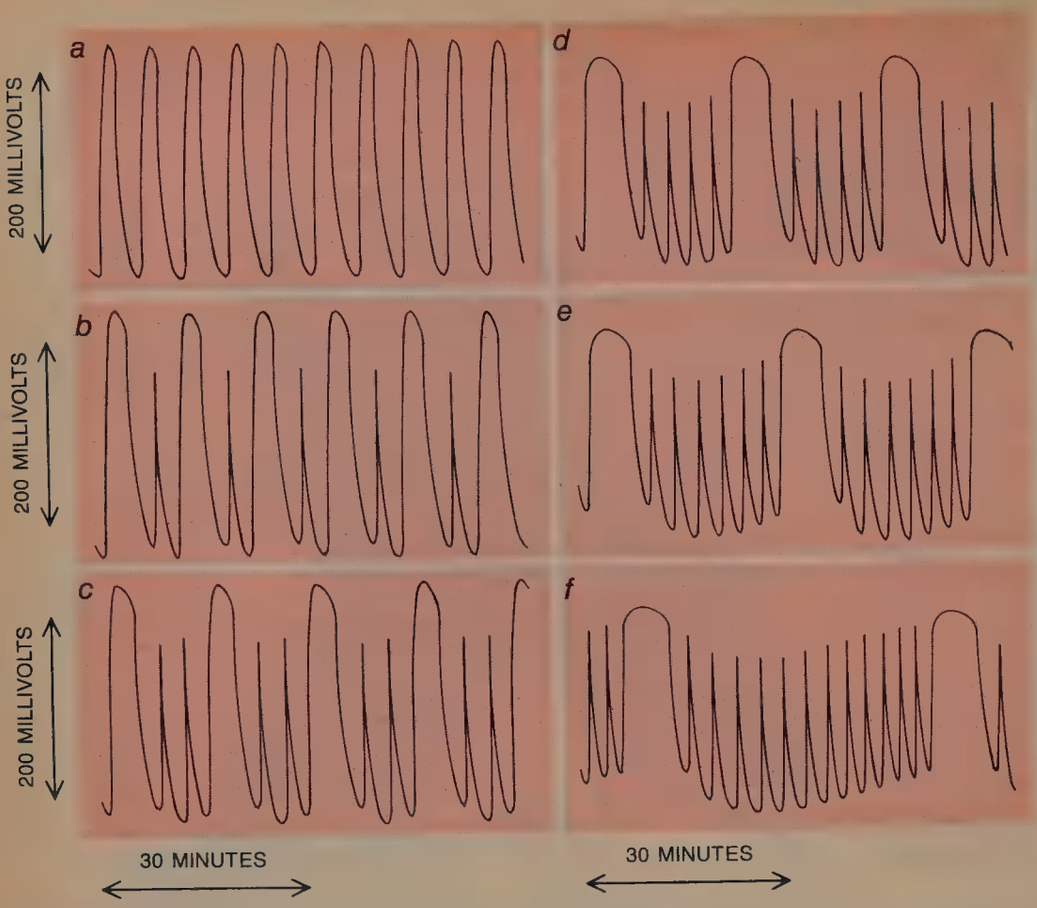
OXFORD UNIVERSITY PRESS
Dept. BN, 200 Madison Avenue, New York, NY 10016

Please send me the following books:
Quantity
___(853907-X) @ $25.00 ___(503267-5) @ $7.95
___(217726-5) @ $25.00 ___(286036-4) @ $8.95

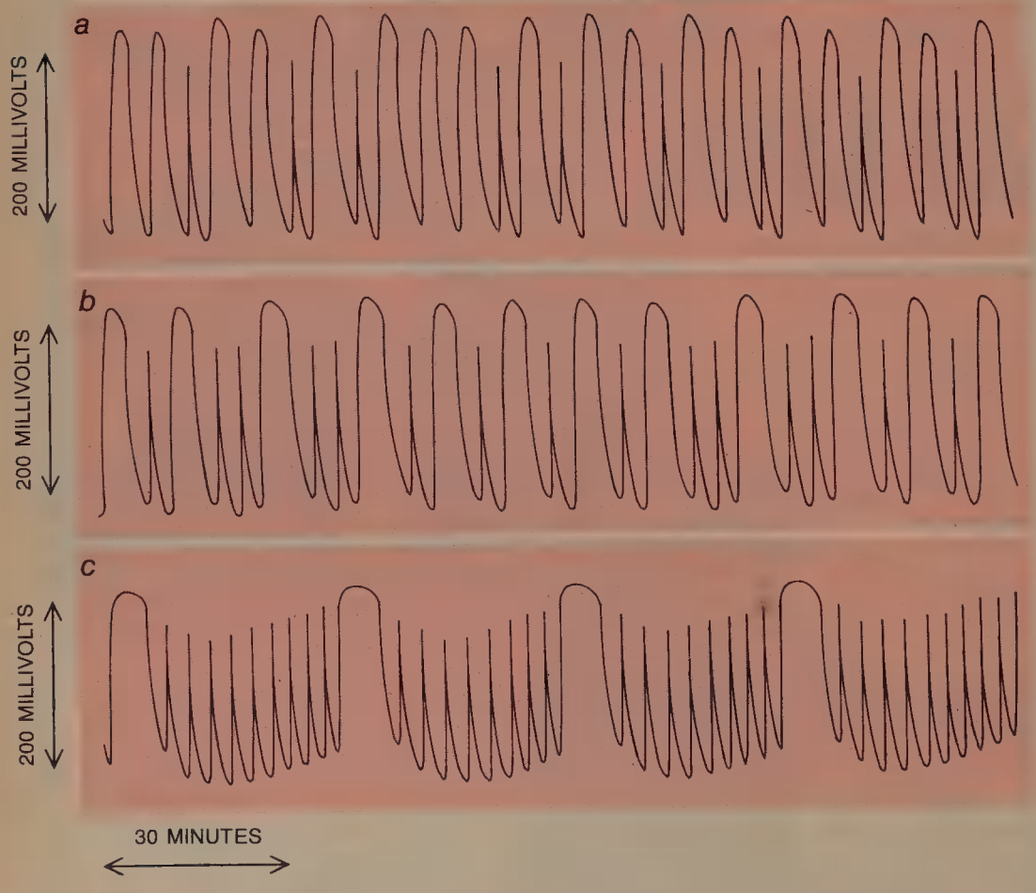
I understand that if I am not satisfied, I may return the purchased books within 15 days for a full refund. Oxford will pay all postage and handling.

☐ I am enclosing check or money order for the full amount.
☐ Charge to my ☐ Visa ☐ MasterCard

Account # _____ Exp. date_____
Signature _____
Print Name _____
Address _____
City _____ State ____ Zip _____



PERIODIC OSCILLATIONS in the chlorite-thiosulfate reaction exhibit a variety of patterns when the input rate to a continuous-flow reactor is altered systematically. Each cycle has one large oscillation and n small ones, where n can vary from 0 to 16 and is the same for all cycles.



APERIODIC, OR CHAOTIC, OSCILLATION is sometimes observed in a narrow range of flow rates between rates that yield complex periodic behavior. One form of chaos consists of an apparently random mixture of two types of periodic oscillation, one with n and one with $n + 1$ small oscillations for each large oscillation. One possible explanation is that uncontrolled variations in the flow rate randomly flip the system back and forth from a rate that gives n-type oscillation to one that gives $(n + 1)$-type. The alternative, which is favored by the authors, is that chaotic oscillation is intrinsic in the dynamics of the chemical reaction at certain flow rates.

existence has aroused considerable controversy in the chemical community, much as the reality of oscillating reactions was questioned a quarter of a century ago, is "chemical chaos." The term refers to the behavior of a reaction in which concentrations neither attain constant values nor oscillate periodically but rather increase and decrease in an apparently random and unpredictable way. We have seen such fluctuations in the BZ system and in one of our new chlorite oscillators, when both are run in a continuous-flow reactor with careful monitoring. They are of particular interest to mathematicians because if they were genuinely inherent in the dynamics of the reaction, they would be physical examples of the mathematical objects known as strange attractors. Briefly and somewhat crudely, strange attractors can turn up when certain equations are iterated, or solved repeatedly, with the result that one iteration serves as the input for the next. In a system of equations with a strange attractor the path mapped out by the successive solutions appears to vary unpredictably from one cycle to the next. (Douglas R. Hofstadter explored the topic in his *Scientific American* department "Metamagical Themas" for November, 1981.)

Skeptical chemists have suggested that chemical chaos may represent nothing more than uncontrolled experimental fluctuations, for example in temperature or flow rate, which randomly push a system from one mode of complex periodic oscillation to another. Although the jury is still out, recent careful experimental and theoretical work by Harry L. Swinney and Jack Turner of the University of Texas at Austin, by Jean Claude Roux, Christian Vidal and their group at the Paul Pascal Research Center and by J. L. Hudson of the University of Virginia all strongly suggest that chemical chaos is a genuine phenomenon inherent in the dynamics of at least some chemical oscillators. The implications of such intrinsic chaos are not yet understood. Might it be, as Henry Adams once observed, that "chaos often breeds life, when order breeds habit"?

Now that chemical oscillators have shown themselves to be not only consistent with the laws of nature but also amenable to systematic design and description, they are coming to play an increasingly prominent role in several areas of science. The switching from one state to another that characterizes an oscillating reaction may hold the key to understanding regulatory processes in living cells, such as the mechanisms that turn on and off the copying of a strand of DNA or the contraction of a muscle. The same forces that create colored rings and layers in unstirred chemical oscillators may also be responsible for the spacing of the rings of Saturn and

the periodic striations found in certain rock formations, which are not readily explained by conventional geologic processes.

It also seems plausible that many and perhaps all of the catalytic reactions fundamental to the chemical industry proceed in an oscillatory manner. The catalyst may alternate between two forms on a time scale too short to be detected by standard analytic procedures. The study of oscillating reactions can provide a testing ground for theories of catalysis and, more generally, for theories of the dynamics of chemical reactions. Any model that can successfully predict the complex behavior of oscillating systems should be able to deal readily with the simpler phenomena observed in less exotic reactions. As more chemical oscillators are found or are deliberately designed, one may hope to obtain key insights into a variety of phenomena, ultimately including those clocks and carousels that operate within us all.

SPATIAL STRUCTURE develops in an oscillator based on chlorite, iodide and malonic acid discovered by the authors. The solution, placed in a shallow glass dish, is initially the uniform purple (*top left*) of a complex of iodine, iodide and a starch indicator. As the reaction proceeds white dots appear, then grow into rings and sets of concentric rings, which annihilate one another on colliding. The last picture was made about 90 seconds after the start of the reaction.

"The abuse of truth ought to be as much punished as the introduction of falsehood."
—Blaise Pascal, *Pensées*

New from Norton—
A correct, comprehensive, and comprehensible reference and introduction to Pascal

Complete with:

- Conformant array parameters
- Readable syntax charts
- Marginal glosses with page references
- Complete BNF
- A tear-out pocket reference card
- Full ISO ANSI/IEEE Standard Pascal (ANSI/IEEE770X3.97–1983)
- And an index!

Standard Pascal User Reference Manual
by DOUG COOPER
Coauthor of the best-selling text, **Oh! Pascal!**

NORTON W.W. Norton & Company, Inc., 500 Fifth Avenue, New York, New York 10110

Please send me_____copies of **Standard Pascal User Reference Manual** (30121-4) at $12.50 each.
Please send me_____copies of **Oh! Pascal!** (95205-3) at $17.95 each.
CA & NY residents please add sales tax.

☐ I enclose my check for $_____.
Charge my ☐ VISA ☐ American Express
☐ MasterCard
Account #_____ Exp. Date_____
Name_____
Address_____
City_____ State_____ Zip_____
SA

Autopsy

Postmortem examination contributes in many ways to medical knowledge and the planning of health care. Yet in recent years the rate of autopsy in the U.S. has declined to only 15 percent

by Stephen A. Geller

The only way to be certain what caused a person's death is to do an autopsy, or postmortem examination, of the body. The autopsy often reveals more than the cause of death. It can uncover diseases that went unrecognized in life, and it is a means of checking on the efficacy of medicine, surgery or other therapy that the person received in life for diseases that had been recognized. Yet barely 15 percent of the deaths in the U.S. are followed by autopsy now, whereas in the years after World War II the autopsy rate was nearly 50 percent. What accounts for the decline? Is the present low rate of autopsy detrimental to the cause of good medical care?

The word autopsy is derived from the Greek for seeing for oneself. As it is usually done, the procedure entails three kinds of seeing: a careful examination of the exterior of the body, a dissection and examination of the major organs and a study under the microscope of tissues from the organs. In the external examination, which includes examination of the orifices of the body, the pathologist is checking for evidence of abnormalities of any kind. They include growths, recent or old wounds and recent or old surgical scars.

Dissection begins with a large incision in the shape of a *Y*. The upper branches of the *Y* start at the front of each shoulder, cross below the nipples and join at the bottom of the sternum (the breastbone). The incision is continued down the middle of the abdomen to the pubis. In the abdomen it extends through the layers of the abdominal wall to enter the peritoneal cavity, which contains many of the major organs including the liver, the kidneys, the stomach, the intestines and the genitourinary system. The tissues are folded to the side to expose the abdominal cavity. Similarly, the sternum is removed to expose the cavities of the chest, which contains the heart, the lungs and the thymus. Each organ and system of organs is examined in place for evidence of abnormalities. As a rule the intestinal tract, from the duodenum to the rectum, is removed before the other organs are dissected.

The dissection of organs is usually done by one of two methods developed in the 19th century by the German physicians Rudolf Virchow and Friedrich Albert von Zenker. In Virchow's method the organs are removed from the body one by one and each organ is dissected and examined separately. Von Zenker's method, derived from a technique developed by his Austrian contemporary Karl von Rokitansky, calls for the organs to be removed from the body en masse and then dissected as organ systems.

The von Zenker approach is best for the examination and preservation of connections, normal or abnormal, between organs and organ systems. One example of the approach is that all the organs, beginning with the trachea and extending to the anorectum, the urethra and (in females) the vagina are removed en bloc. Once they are out the organs of the chest are separated from the abdominal and pelvic organs. The heart and the major blood vessels associated with it are examined in relation to the respiratory system and then are dissected free. When the abdominal organs are studied, the liver, the pancreas, the gall bladder and the bile duct are examined in terms of their relation to one another and to the stomach, the small intestine, the spleen and the associated blood vessels before they are separated and dissected individually.

The brain can be examined by cutting into the back of the scalp and part of the skull. It is studied in position, freed from its attachments to blood vessels and nerves and removed (by severing the lower part of the medulla oblongata). The spinal cord is removed by freeing it from its lateral attachments and extracting it through the skull or by cutting through the vertebral bodies.

In these steps as in other parts of the autopsy the method can be modified to meet the needs of the clinical problem to be solved. The breasts and the muscles can be studied without additional incisions of the skin. The face and the limbs are usually not dissected except with special permission in cases where the information obtained may be of particular clinical significance. All the incisions are of the surgical type and can be covered with clothing for a funeral.

After the organs have been studied as systems they are examined individually. Weights and measurements are recorded for comparison with standard values. Incisions are made in each organ to identify fine details. Grossly observable features are recorded in a standardized terminology.

For examination by microscope thin pieces of tissue from various organs are embedded in wax. Extremely thin slices of tissue are cut from these paraffin blocks, placed on glass slides, stained with various dyes and examined with the light microscope. For special problems an electron microscope may be brought into play. Additional techniques may include attempts to recover microorganisms that may have contributed to disease and chemical analyses designed to detect a wide range of substances perhaps made internally during life or taken in from outside.

The entire procedure takes from two to four hours, depending on whether the autopsy is total or limited, meaning that it deals with only one organ or a few organs that are thought to be of particular relevance to the case. Probably about 75 percent of the autopsies done in the U.S. are total. When the autopsy is finished, the organs are returned to the body (except for those that by permission are being donated to living people or are serving as a source of pharmacologic substances such as human growth hormone) and the incisions are sewn up. The final step is the compilation of a report of the findings.

The autopsy has been an integral part of medical practice for more than 100 years, but its purposes have changed somewhat. At one time the autopsy's main role was to identify the basis of

disease and the cause of death, and often that is still the chief reason for doing one. When someone dies unattended or without medical care, the autopsy is the only means of understanding the terminal events.

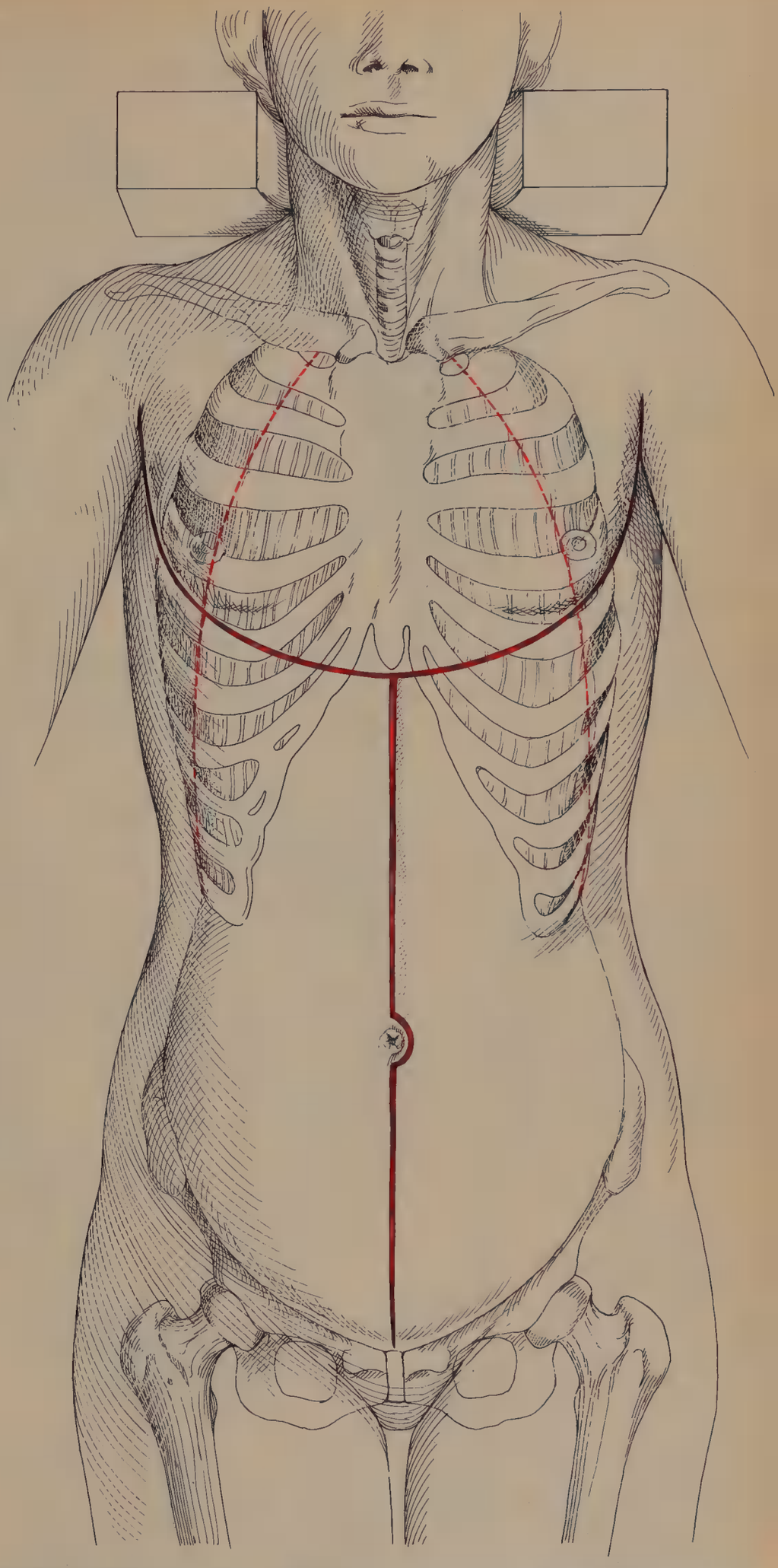
In most of the larger cities of the U.S. a medical examiner has the responsibility for reviewing the deaths that take place outside the hospital. Almost without exception, however, the office of the medical examiner is overburdened with deaths related to possible violations of law, and people whose death seems to have been due to natural causes are not autopsied.

For many of the deaths that occur in a hospital the principal disease has been diagnosed before death. In these cases the autopsy serves the particularly important function of a quality-control device. It monitors the accuracy of interpretation of diagnostic tests, determines the efficacy and toxicity of therapeutic agents, allows for the detection of conditions that may have been important but were either clinically inapparent or obscured by the most prominent disease and helps to monitor the influence of environmental factors on physiological processes. In addition the autopsy is a highly effective method for the education of medical students and the continuing education of physicians on the variations in human disease.

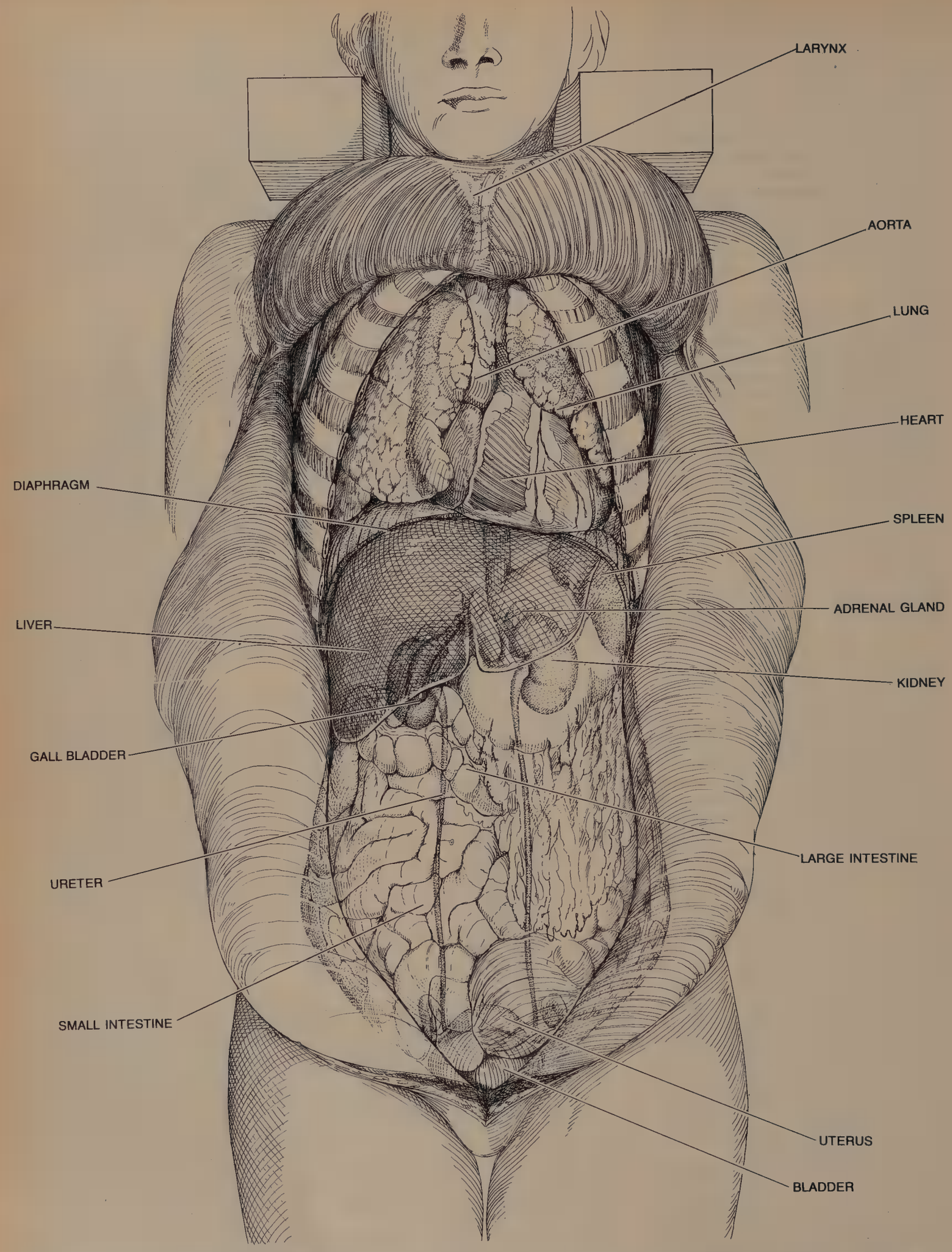
A measure of the value of autopsies is that even with the many recent advances in medicine the major diagnosis turns out to have been wrong in as many as 40 percent of the people autopsied. Even in the best medical centers the major disease is shown to have been incorrectly diagnosed in more than 15 percent of the autopsied cases and the immediate cause of death to have been clinically unrecognized in more than 40 percent.

Since most patients who die are not examined after death, there are no reliable figures on the overall accuracy of medical diagnosis. One indication that the rate of error is consistently high even in major medical centers was provided by the late Edward A. Gall, who was professor of pathology and chairman of the department of pathology of the University of Cincinnati School of Medicine. He documented the remarkable similarity of the percentage of errors in diagnosis, as disclosed by autopsy, at two major institutions over the 50-year period from 1912 to 1962.

A clear corollary of this situation is that the health statistics dealing with the causes of death are not useful. Indeed, the inaccuracy of death certificates has been well documented. Linda W. Engel and her co-workers at the National Cancer Institute assessed the accuracy of certification of the underlying cause of death in 2,557 cases autopsied in vari-



FIRST INCISION in the standard autopsy procedure is made following an external examination of the body for abnormalities. The incision (*color*) is in the form of a *Y* and is made to enable the pathologist to examine the organs of the chest, the abdomen and the pelvis. Then (*broken line*) the ribs are cut so that the breastbone can be removed to expose the chest cavities.

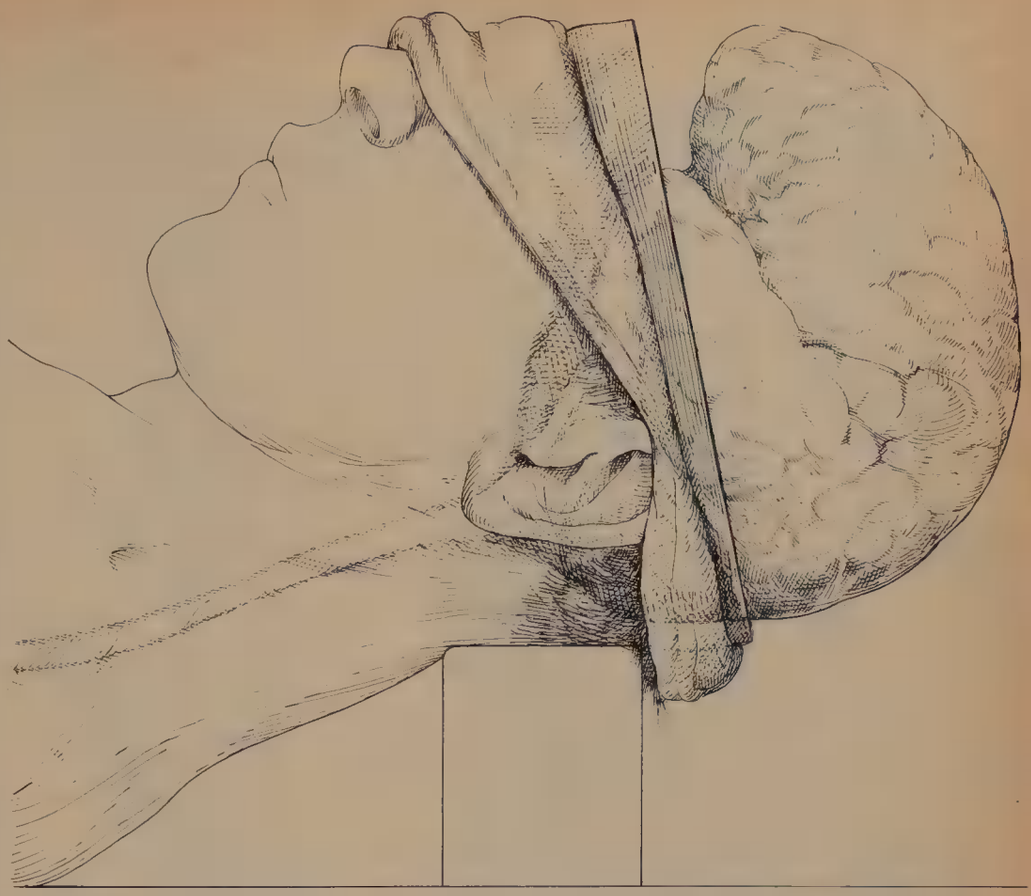


EVISCERATION OF ORGANS is done according to one of two different methods. Here the method is to remove the organs en masse and examine and dissect them as organ systems before dissecting the organs individually; the method was developed in the 19th century by the German physician Friedrich Albert von Zenker. In the other method, devised by Rudolf Virchow, the organs are removed from the body one by one and dissected separately. After the autopsy the organs are returned to the body and the incisions are sewn up.

ous hospitals. As is common practice in many institutions, each death certificate was filled out by one of the physicians responsible for the care, without benefit of the autopsy findings. Engel found that the underlying cause of death had been improperly or inaccurately recorded in 42 percent of the cases.

Since only 15 percent of the two million people who die in the U.S. each year are autopsied and the death certificates are often incorrect in even that small number of cases, it is reasonable to assume that each year at least a million death certificates are in error. How much of the nation's health-care planning is based on those certificates? Only postmortem examination can provide the basis for accurate statistics.

Similarly, only the systematic performance of autopsies can provide the data necessary for understanding the diseases that affect a nation's population. At present the pattern of autopsies is skewed. Of the 69 categories for cause of death listed in *Vital Statistics of the United States, 1977,* only six had autopsies performed in more than 50 percent of the cases: homicide, abortion, other complications of pregnancy, meningococcal infections, bacillary dysentery and amebiasis and "other external causes." These six conditions account for fewer than 2 percent of the deaths each year and are hardly the nation's most pressing health problems. In contrast only 9.9 percent of the more than 950,000 deaths due to major cardiovascular disease and 11.3 percent of the more than 385,000 deaths due to malignancy were autopsied.



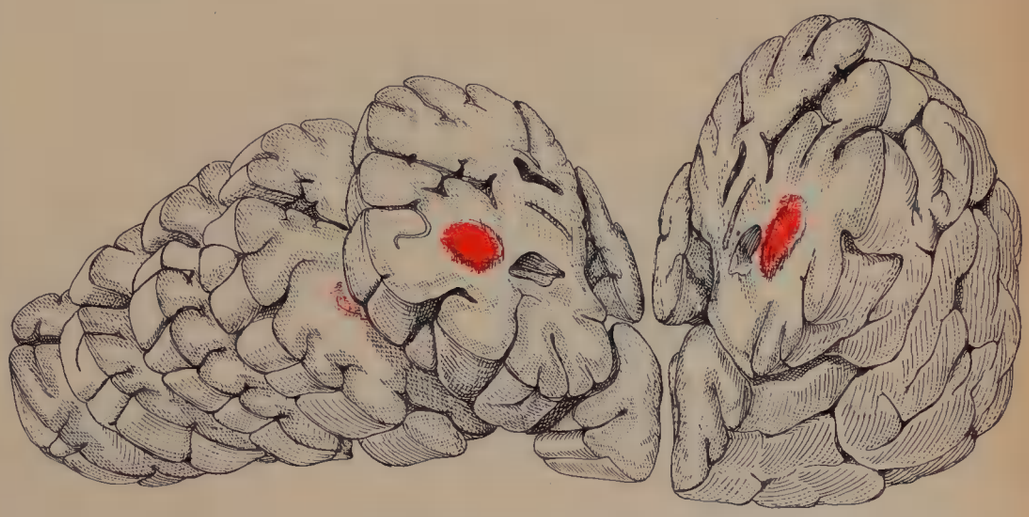
EXAMINATION OF THE BRAIN begins with a study of it in its normal position. The brain is then removed for closer examination. The spinal cord can also be removed through the skull.

Many other benefits of the autopsy can be cited. It is a vital complement to the considerable number of new diagnostic techniques that have been introduced during the past decade, such as computer-assisted tomography (the CAT scan), enabling the diagnostician to check the validity of the interpretations that have been made on the basis of readings from the machines. The autopsy is an excellent means of gauging the effectiveness of medications and, equally important, their relative toxicity, particularly for the many drugs employed in treating cancer. Those drugs often exert a strong effect on normal cells as well as on malignant ones. By applying the classic approach of observing the patient during life and correlating the findings at death with the clinical events the mechanisms of such toxic phenomena become known. For example, the tendency of Adriamycin, a potent anticancer agent, to cause heart failure became understood in this way, and the dosage was modified to minimize the likelihood of fatal injury.

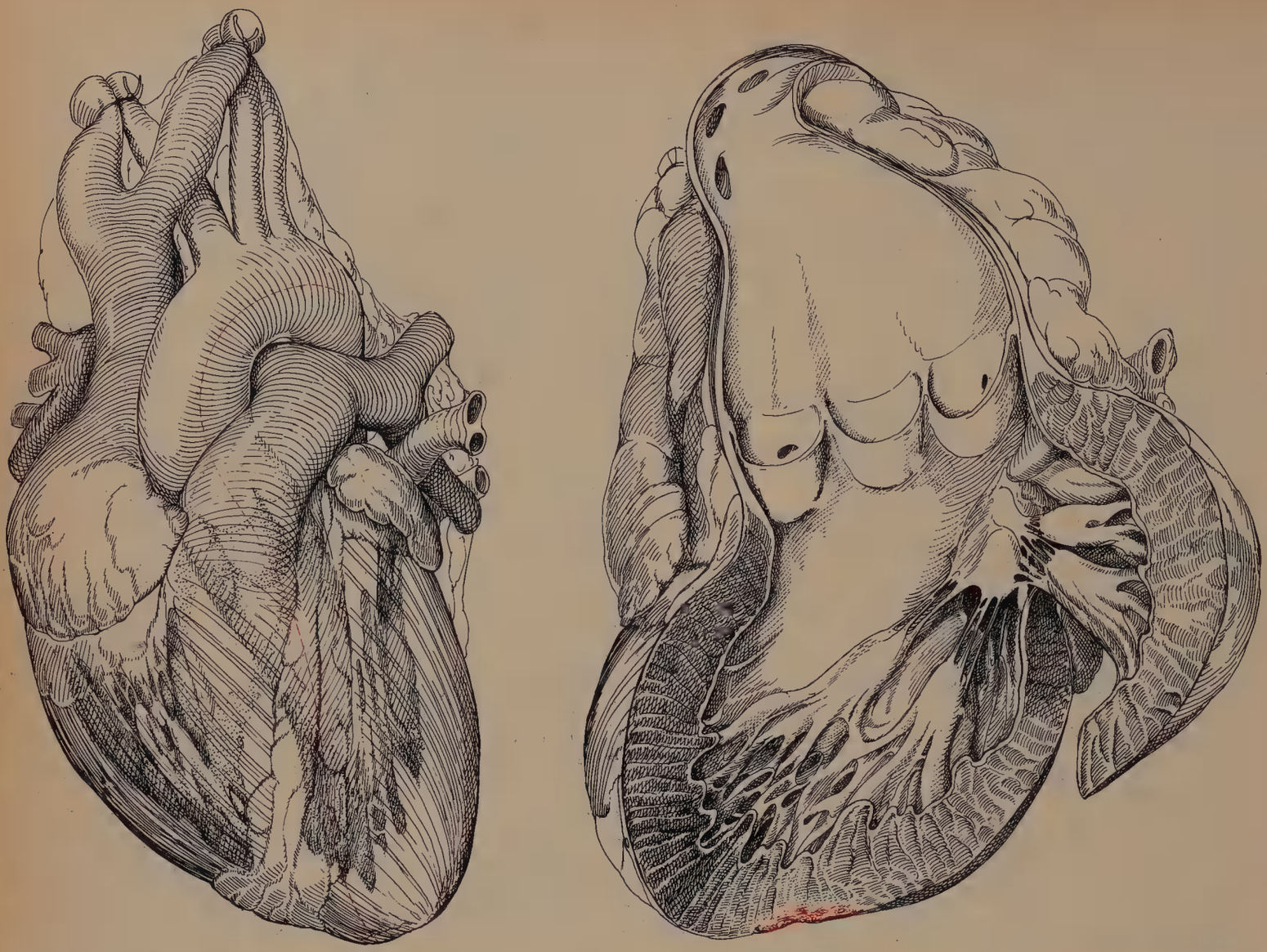
What are the long-range effects of prosthetic devices on tissues? In the early 1960's, when prosthetic heart valves were first being widely adopted, autopsies yielded the practical information on size, shape and materials that allowed for improvements in the design of the devices and for the refinement of the criteria employed to choose patients who were most likely to be helped by the artificial valves.

The autopsy is a singularly effective tool for recognizing new disorders that may be due to occupational or environmental influences. In the past decade autopsies helped to elucidate the basis of the development of a rare, highly malignant tumor—angiosarcoma of the liver—affecting workers in factories producing vinyl chloride. The repeated recognition of the tumor at autopsy was eventually followed by the appropriate epidemiologic and experimental studies, which led ultimately to the control of the conditions that gave rise to the tumor. The autopsy also helped to delineate the conditions associated with occupational exposure to asbestos.

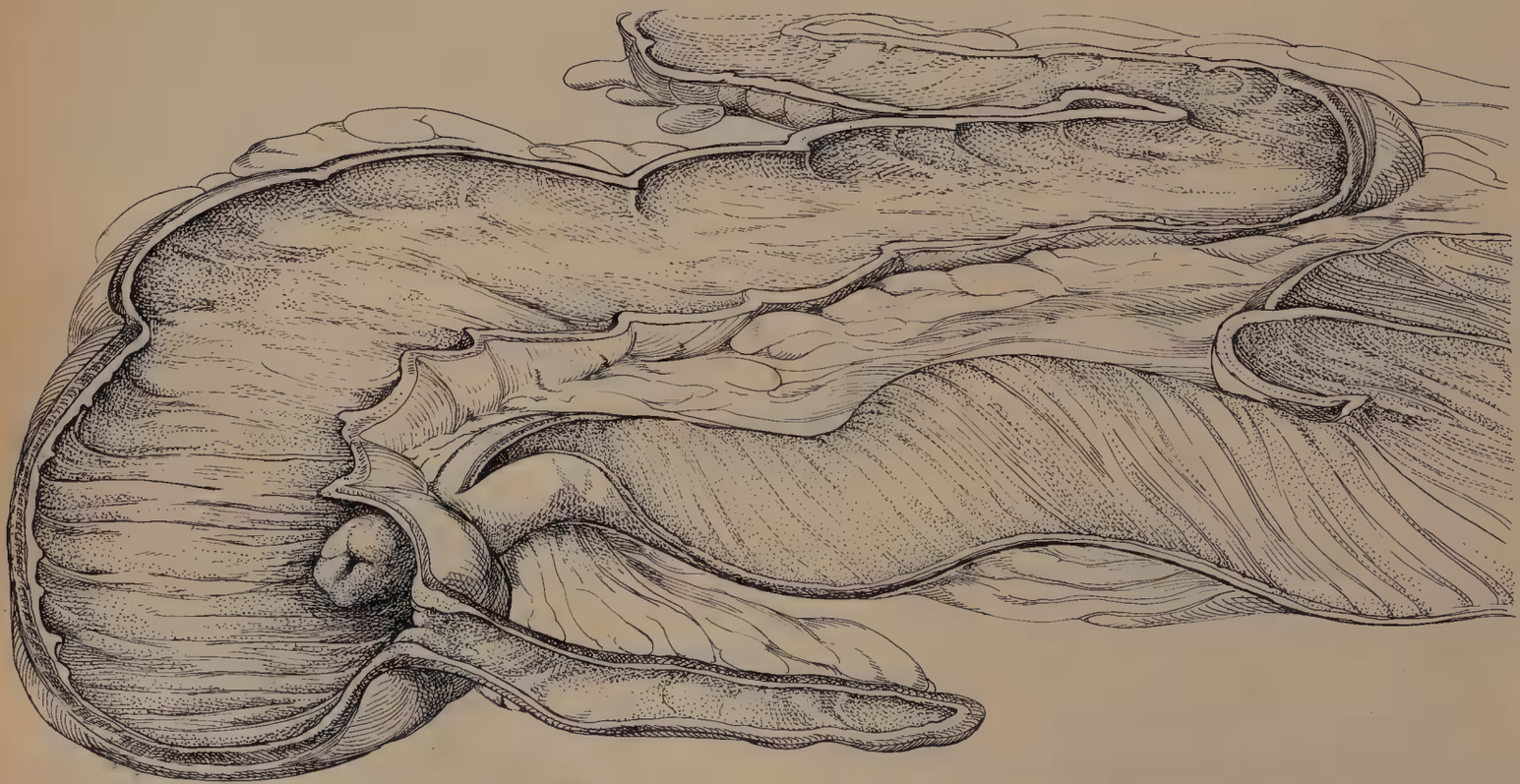
Environmental pathology is one of the newest, most publicized and least defined areas of study in medicine. The bulk of the evidence associating specific environmental factors with effects



NEXT STAGE in the examination of the brain is to make incisions in it in order to examine details and look for indications of disease. Here a lesion (*color*) is found; it is a tumor of the brain.



STUDY OF THE HEART, once it has been examined in relation to its associated blood vessels and the lungs, entails an incision (*broken line*) to open a ventricle, in this case the left one. When the ventricle has been opened (*right*), it is examined for evidence of disease. Here an infarct (*color*), an area where tissue has died when the circulation of blood was cut off by a clot, is seen in the bottom wall of the ventricle.



INTESTINE is examined separately after it has been studied in relation to the stomach and other associated organs, including the pancreas, the gall bladder and the bile duct. Here following dissection and an incision for closer inspection the intestine is found to be normal.

on organs remains circumstantial. The enormous potential of the autopsy to establish such relations is almost entirely unrealized. Some associations, such as air pollution and lung disease, are likely to be found. How many others can be identified and correlated with specific causative factors? What are the early, preclinical manifestations in the tissues of environmental injuries?

Organs and individual cells obtained at autopsy have been successfully maintained in culture for prolonged periods of time, making possible many in vitro experiments on human tissue. Human tissues recovered at autopsy can also be maintained in immunologically "nude" mice for many months, allowing for a multitude of in vivo experiments.

It is in the light of considerations such as these that one wonders why the rate of autopsy has declined so significantly over the past 30 years. A prominent reason is the feeling among many physicians that diagnoses are almost always established during life and that nothing new is learned from the autopsy. The statistics cited above on errors of diagnosis, together with the long list of diseases discovered or elucidated at autopsy in recent years (among them Legionnaires' disease and the toxic-shock syndrome), suggest that the opinion is insupportable.

A more important factor may be the reluctance of the physician to accept death, regarding it as a failure to be quickly forgotten. A compounding factor is that most patients who die in a hospital have been cared for by more than one physician. The trend toward specialization by physicians and the decrease in the number of general practitioners have led to a relative lack of familiarity between physician and patient. Doctors are no longer confidants and lifelong friends, and so they often lack the determination and confidence to pursue the matter of getting permission for an autopsy from the family at the time of death.

It is also possible that autopsy has declined because of a concern among physicians that it will disclose an error in diagnosis or treatment and an associated concern about malpractice suits. Although the data on this point are skimpy, documented instances in which the performance of an autopsy has been a major contributing factor in a malpractice suit are exceedingly rare. In fact, a thorough postmortem examination may support the defense of a physician whose medical care is shown by the examination to have been consistent with accepted practice.

Often the family of the person who has died is reluctant to give permission for an autopsy. The assertion "He's suffered enough" frequently goes unanswered, although the family member

INVEST YOURSELF

A windmill to pump water for "salt farming" in India. More efficient woodburning stoves for the Sahel. Photovoltaic irrigation pumps for the Somali refugee camps.

All these are solutions to technical problems in developing countries. Devising such solutions is no simple task. To apply the most advanced results of modern science to the problems of developing areas in a form that can be adopted by the people requires the skills of the best scientists, engineers, farmers, businessmen—people whose jobs may involve creating solid state systems or farming 1000 acres, but who can also design a solar still appropriate to Mauritania or an acacia-fueled methane digester for Nicaragua.

Such are the professionals who volunteer their spare time to Volunteers in Technical Assistance (VITA), a 20 year old private, non-profit organization dedicated to helping solve development problems for people world-wide.

Four thousand VITA Volunteers from 82 countries donate their expertise and time to respond to the over 2500 inquiries received annually. Volunteers also review technical documents, assist in writing VITA's publications and bulletins, serve on technical panels, and undertake short-term consultancies.

Past volunteer responses have resulted in new designs for solar hot water heaters and grain dryers, low-cost housing, the windmill shown above and many others. Join us in the challenge of developing even more innovative technologies for the future.

VITA **Putting Resources to Work for People**

3706 Rhode Island Avenue, Mt. Rainier, Maryland 20712 USA

Foresight is better than hindsight.

Have our leaders lost sight of the real issues involved in matters of nuclear strategy? Caught in a tangled web of political momentum, economic concerns, and their own organizational pressures, a small political and military elite is now responsible for critical strategic decisions determining the fate of civilization itself.

We cannot afford to leave such irrevocable decisions to the "experts." But it is only through gaining an understanding of the primary political issues and weapons systems that we can hope to make our voices heard and have any effect on the course of future events. This understanding is within our grasp—and the future is the responsibility of us all. **Knowledge is our best defense.**

THE PRISONERS OF INSECURITY
Nuclear Deterrence, the Arms Race, and Arms Control
Bruce Russett

Demystifying the strategic nuclear debate by clarifying the fundamental political issues and providing the facts and figures necessary to an informed opinion, *The Prisoners of Insecurity* was written to help civilians think for themselves about the complex issues involved and demonstrates that most of the basic questions about national security and arms control are political rather than technological.

1983, 204 pages
hardbound: 1471-X $14.95
paper: 1472-8 $7.95

WINDING DOWN
The Price of Defense
The Boston Study Group

This realistic assessment of the defense capability, needs, and budget of the United States argues that our military expenditures can be reduced significantly without sacrificing a safe and effective defense program. *Winding Down* was originally published as *The Price of Defense* and is now being reissued in response to the reorganization of the U.S. budget and increased sums allocated for national defense.

1982, 359 pages
paper: 1498-1 $7.95

LAST AID
The Medical Dimensions of Nuclear War
International Physicians for the Prevention of Nuclear War

A collection of articles written by physicians from around the world, *Last Aid* considers the physical and psychological tolls of a nuclear holocaust. This important study shows that medical professionals could not begin to cope with the problems that would ensue in the aftermath of a nuclear war, and contends that thermonuclear warfare is a lethal menace to civilization and the human species.

1982, 338 pages
hardbound: 1434-5 $19.95
paper: 1435-3 $9.95

These titles are available at fine bookstores and may also be purchased directly from the publisher. Checks made payable to W. H. Freeman and Company should be sent to our California office (please include $1.50 for postage and handling; California, New York, and Utah residents please add appropriate sales tax).
Promotion Department, W. H. Freeman and Company, 660 Market Street, San Francisco, CA 94104

Medicine is always changing, now there's a text that changes right along with it.

SCIENTIFIC AMERICAN *Medicine*.

The answers that you get from SCIENTIFIC AMERICAN *Medicine* are current, because each month new chapters replace the old. You will know where to look for the answers, because this advanced text follows the familiar structure of medical knowledge, changing as medicine changes.

Comprehensive. Leading authorities from Harvard and Stanford cover the full range of medicine for you the seasoned clinician.

Complimentary CME. At no extra cost, we offer a CME program in which you can earn 32 top credits* per year.

Trial Offer. We invite you to try SCIENTIFIC AMERICAN *Medicine* for two months with our compliments. Send us the coupon and you will receive the two-volume text and two monthly updates. You may also take a CME test for credit. At the end of 60 days, if you decide to continue the subscription, you will owe us $220 for the first 12 months (current renewal is $170); otherwise, return the books; there is no charge.

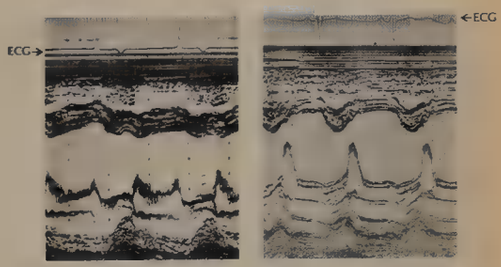
Please mail the coupon today so that we can put medical progress to work for you.

You may also order your trial copy by calling toll-free 1-800-345-8112 (in Pennsylvania, 1-800-662-2444).

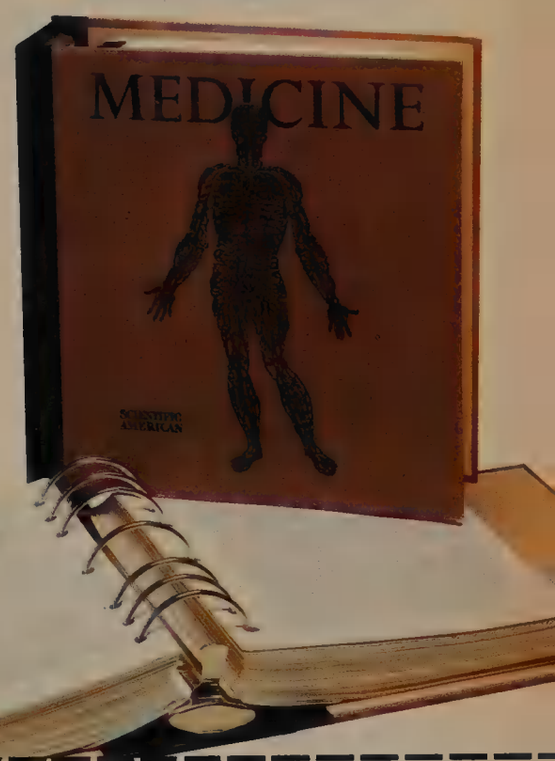
*As an organization accredited for continuing medical education, the Stanford University School of Medicine designates this continuing medical education activity as meeting the criteria for 32 credit hours in Category 1 for Educational Materials for the Physician's Recognition Award of the American Medical Association, provided it has been completed according to instructions.

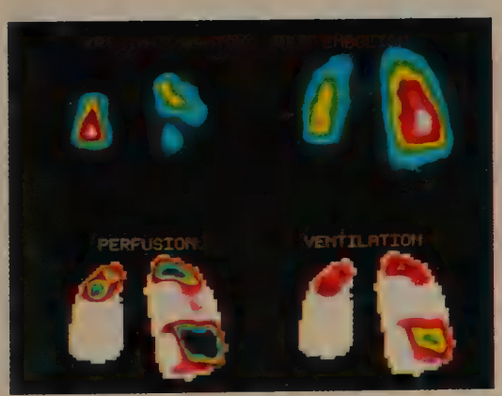
This program has been reviewed and is acceptable for 32 prescribed hours by the American Academy of Family Physicians.

This program has been approved by the American College of Emergency Physicians for 32 hours of ACEP Category 1 credit.

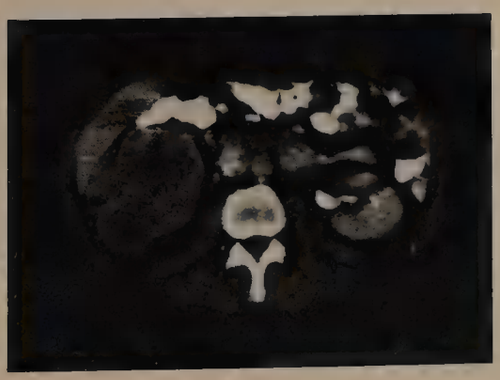


Echocardiograms from patients with aortic regurgitation.





Computerized scintigraphy reveals pulmonary thromboembolism.



Abdominal computed tomogram reveals large renal carcinoma replacing part of right kidney.

SCIENTIFIC AMERICAN MEDICINE

415 Madison Avenue, New York, New York 10017

Please enroll me as a subscriber to SCIENTIFIC AMERICAN *Medicine*. On receipt of this coupon you will send me the advanced two-volume text described in your announcement and update it regularly by sending me new monthly subsections. I understand that the annual subscription of $220 for SCIENTIFIC AMERICAN *Medicine* is tax deductible. If I am not entirely satisfied, I may cancel at any time during the first 60 days, returning all materials for a *complete refund*.

☐ Please enter my subscription for SCIENTIFIC AMERICAN *Medicine*
☐ I shall also enroll in the CME program
☐ I enclose a check made out to SCIENTIFIC AMERICAN *Medicine* for $220*
☐ Bill me ☐ VISA ☐ MasterCard
Expiration Date_____ Account Number_____
*Please add sales tax for California, Illinois, Massachusetts, Michigan and New York

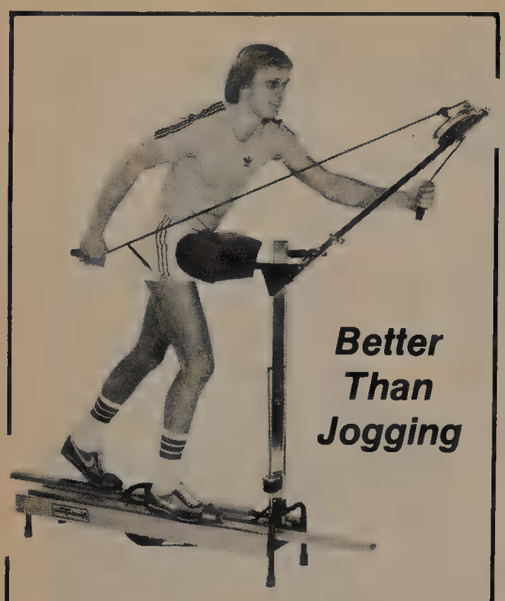
Name_____ MD Specialty_____

Address_____

City_____ State_____ Zip_____

Signature_____

Please allow six to eight weeks for delivery. All payments must be in U.S. dollars. CME is available outside the U.S. and its possessions for an additional $65. Surface routes used unless airmail delivery is prepaid.



NordicTrack

Jarless Total Body Cardiovascular Exerciser Duplicates X-C Skiing for the Best Motion in Fitness

The enjoyable sport of cross-country skiing is often cited by physiologists as the most perfect form of cardiovascular exercise for both men and women. Its smooth, fluid, total body motion uniformly exercises more muscles and higher heart rates seem easier to attain than when jogging or cycling. NordicTrack closely simulates the pleasant X-C skiing motion and provides the same cardiovascular endurance-building benefits—right in the convenience of your home. Makes a year round, consistent exercise program easily attainable. Eliminates the usual barriers of time, weather, chance of injury, etc. Also highly effective for weight control.

More Complete Than Running
NordicTrack gives you a more complete work out—conditions both upper body and lower body muscles at the same time. Fluid, jarless motion does not cause joint or back problems as jogging or running often does.

More Effective Than Exercise Bikes
NordicTrack's stand-up skiing motion more uniformly exercises the large leg muscles and also adds important upper body exercise. Higher pulse rates, necessary for building fitness, seem easier to attain because the work is shared by more muscle mass.

Even Better Than Swimming
NordicTrack more effectively exercises the largest muscles in the body, those located in the legs and buttocks. When swimming, the body is supported by the water, thus preventing these muscles from being effectively exercised. The stand up exercising position on the NordicTrack much more effectively exercises these muscles.

Practical, Takes Little Storage Space
Folds compactly in only 30 seconds. Requires only 15 by 17 inches storage area.

A Proven, High Quality Durable Product
NordicTrack is in its 7th year of production. NordicTrack is quiet, motorless and has separately adjustable arm and leg resistances. We manufacture and sell direct. One year warrantee, 15 day trial period with return privilege.

Call or write for free brochure
TOLL FREE 1-800-328-5888
8 AM-5 PM MON.-FRI.
PSI 124F Columbia Crt. Chaska, MN 55318
Minnesota 612-448-6987

may actually be asking for reassurance that any suffering is over. Some families fear mutilation of the body, although this objection can usually be overcome by a simple description of the procedures. A family that has given permission for an autopsy in the past may not have been told of the results and so may see little reason for acceding again. Religious objections may be raised, although there are no edicts against autopsy by the Catholic church and the objections raised by orthodox Jews are based on rabbinical interpretations of the 18th century. In some communities undertakers discourage families from giving permission for autopsy because they do not want their own work delayed, even though the autopsy takes only a few hours.

Another factor in the decline of autopsy is that physicians have not had occasion to become interested in the subject. Pathology in general and autopsy in particular have been relatively deemphasized in medical schools. Participation in at least one autopsy was once a requirement for every medical student, but in many medical schools today it is not even an option.

In many hospitals the autopsy room is far removed from the areas where the living patients are. Clinicians think they do not have time to attend the autopsy,

VIEW OF THE LUNG after it has been removed and partially dissected reveals ■ disease (*color*) in the form of a tumor invading the bronchus, which is ■ subdivision of the trachea and carries air into the lung. This is the right lung; ■ similar examination would be made of the left one.

Osborne™ brings you the comparison IBM® and Apple® don't want you to see.

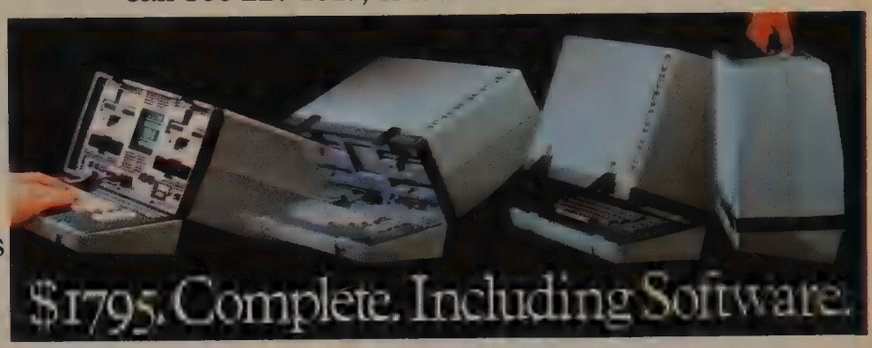
Other computer companies dazzle buyers with an array of options and add-ons that makes the final price hard to determine and makes the computer hard to buy, complex to assemble, and very difficult to carry.

We believe in making personal computers that are easy to learn and use. And that starts with making computers easy to *buy*.

The Osborne 1™ Personal Business Computer. One simple price, $1795, buys it all.

And it all comes in a portable case you can take with you wherever you work. Because once you go to work with an Osborne, you won't want to work any other way.

For your nearest dealer, call (in California) 800 772-3545, ext. 905; (outside California) call 800 227-1617, ext. 905.



OSBORNE™
COMPUTER CORPORATION

	OSBORNE 1™	IBM PERSONAL®	APPLE II®
Computer with 64K RAM, two floppy drives[A], keyboard and CRT:	$1795	$3240[B]	$3120[C]
Serial communications:	INCLUDED	EXTRA COST	EXTRA COST
Modem Connection:	INCLUDED	EXTRA COST	EXTRA COST
IEEE 488 Instrument communications:	INCLUDED	EXTRA COST	EXTRA COST
BASIC interpreter[D]:	INCLUDED	INCLUDED	INCLUDED
Business BASIC[E]:	INCLUDED	EXTRA COST	EXTRA COST
CP/M® Control Program:	INCLUDED	EXTRA COST	F (see below)
Word Processing[G]:	INCLUDED	EXTRA COST	EXTRA COST
Electronic Spreadsheet[H]:	INCLUDED	EXTRA COST	EXTRA COST
Carrying Case:	INCLUDED	EXTRA COST	EXTRA COST
TOTAL PRICE[I]:	**$1795**	**$4000–4700**	**$4000–4700**

A. The Osborne 1™ includes two built-in 100K byte floppy disk drives. The IBM® and APPLE II® drives provide approximately 160K bytes of storage. **B.** From the IBM Product Center Personal Computer Price Schedule. **C.** From the Apple Computer Suggested Retail Price List. **D.** The Osborne includes MBASIC® from Microsoft. **E.** The Osborne includes CBASIC®, a business-oriented BASIC language from Digital Research.™ **F.** The Osborne includes CP/M®, the industry-standard control program from Digital Research. The list of software packages which will run with CP/M is considerable. IBM offers CP/M 86 (a version of CP/M) at extra cost. There are optional hardware systems which allow the Apple II to run CP/M; the Apple II control program is highly comparable to CP/M. **G.** The Osborne includes WORDSTAR® word processing with MAILMERGE®—products of MicroPro™ International. **H.** The Osborne includes SUPERCALC™, the electronic spreadsheet system from Sorcim Corporation. **I.** Exact price comparisons cannot be presented, because the software and hardware options chosen to create the "equivalent" of the Osborne 1 Personal Business Computer vary in price. The range indicated was computed using price lists from IBM and Apple. Documentation of the computations are available on request from Osborne Computer Corporation. Trademarks: OSBORNE 1: Osborne Computer Corporation; SUPERCALC: Sorcim Corporation; Digital Research: Digital Research, Inc.; Registered Trademarks: WORDSTAR, MAILMERGE: MicroPro International Corporation of San Rafael, CA; MBASIC: Microsoft; CBASIC, CP/M: Digital Research, Inc.; IBM: IBM Corporation; Apple, Apple II: Apple Computer Corporation.

IT CAN ACTUALLY MAKE 55 M.P.H. INTERESTING.

If you think that's impossible, you've never driven the Volvo Turbo.

A car whose handling equipment can turn a curve in the road or a trip to the supermarket into a driving adventure.

Its turbocharged 4-cylinder engine can dust a V-8 off the line. Automotive writers have described it as "A blast." "Spectacular." "Like cutting in an afterburner."

Maybe you think speed limits, emissions controls and government mileage requirements have made driving humdrum. But that's only because you don't own an interesting car.

THE TURBO

By Volvo.

which often coincides with other activities in the hospital or with office hours. Even when an autopsy has been done, the results may not be communicated clearly and quickly to the responsible doctor. As William C. Roberts, chief of pathology at the National Heart, Lung, and Blood Institute, has pointed out, the pathology department in a hospital has many responsibilities, virtually all of them viewed as being more pressing than autopsies, so that "the autopsy reports are typed last by the secretaries, the paraffin blocks on autopsies are cut last by the technicians and the staff pathologists delay 'the signing out' of autopsies to the last because of more important duties." This almost inevitable delay further discourages physicians from seeking permission for autopsies.

Disenchantment also follows because of the disparity between the skills of the clinician and those of the person performing the autopsy, which is generally relegated to the most junior member of the pathology department. That person lacks the knowledge, judgment and experience to effectively translate the morphologic observations into clinically relevant comments. The autopsy is second to surgical pathology and the clinical laboratory, both of which serve the living directly, and these are the places where the senior members of the staff are most active.

Over the past 20 years cell biology has become more and more a part of the pathologist's responsibility and interest, the autopsy less and less. Senior pathologists have turned increasingly to sophisticated scientific or diagnostic endeavors. In a teaching institution the autopsy cannot compete with the exciting areas of surgical or experimental pathology as a basis for professional recognition and advancement. Senior pathologists seek other areas of academic expression and so become unavailable for immediate consultations with clinicians who need autopsy interpretations. The junior pathologist is deprived of any opportunity to learn from an instructive two-way discussion of clinical problems and concludes that the autopsy room is not the place to find success.

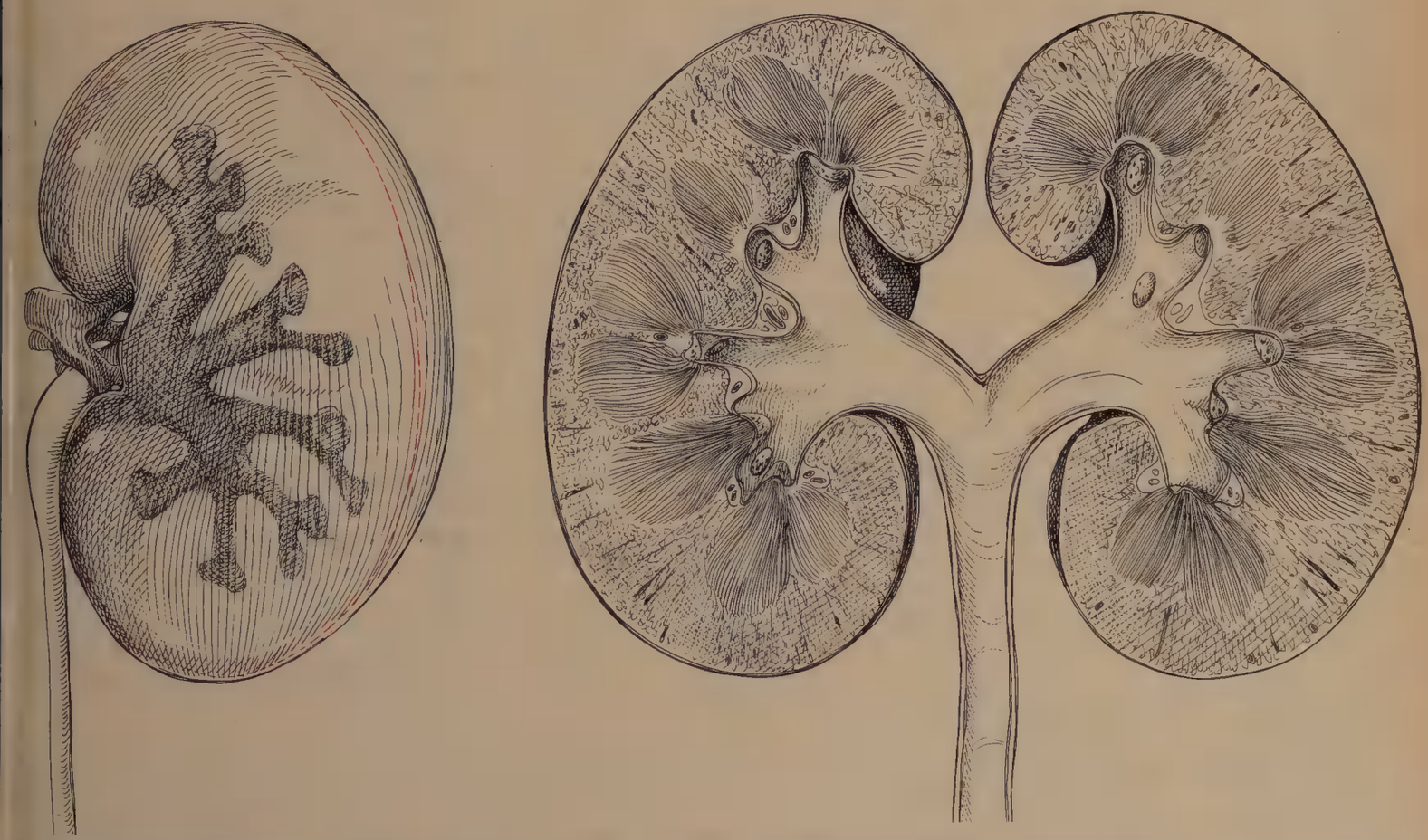
This unsatisfactory situation is further compounded when the autopsy is done by an autopsy technician or a pathologist's assistant rather than by a physician. Assistants of this kind have been utilized recently as a way of releasing the pathologist for ostensibly more important activities. The practice is also thought to have financial advantages. The trend may be shortsighted. If autopsy is to remain a significant component of the practice of medicine, it will require more of the pathologist's time and interest rather than less.

What can be done to reverse the decline of autopsy? To a great extent the pathologist is the key. The pathologist and the clinician must gain a renewed understanding of the critical role autopsy can play in health care. Hospital administrators must be made to understand the major service autopsy can provide for evaluating the cost-effectiveness of hospital activities and for maintaining excellence in medical care.

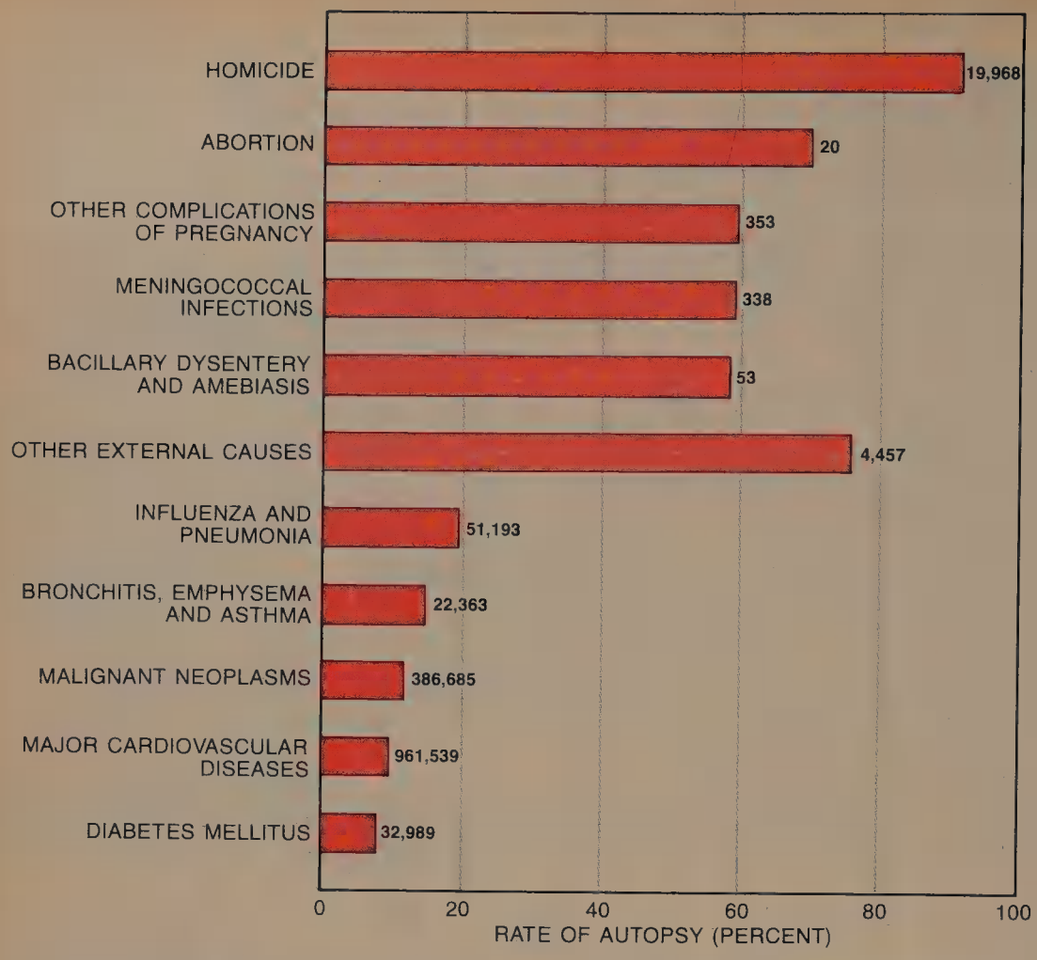
New approaches to the information-handling aspects of the autopsy would be helpful, but they cannot provide a remedy for the pathologist's failure to exercise a dynamic role. Before an autopsy is begun the pathologist and the clinician must identify together the major clinical problems that arose with the patient and the forms of therapy provided. After the autopsy the morphologic findings must be immediately presented to the physician to maximize their instructional value.

Families must be told of the results of the autopsy, and the significance of the key findings must be explained to them. A copy of the final report on the autopsy, with appropriate interpretations, should be given to the closest relative of the deceased. In addition the importance of autopsies should be more fully explained to the public. Hospitals should include a brief discussion of the advantages of the autopsy in the information booklets they give patients, just as they describe X rays, blood tests and many of their other services. Physicians as well as the public need to be told about the many recent contributions of autopsy to medical knowledge and about its vital role in monitoring the quality of medical care.

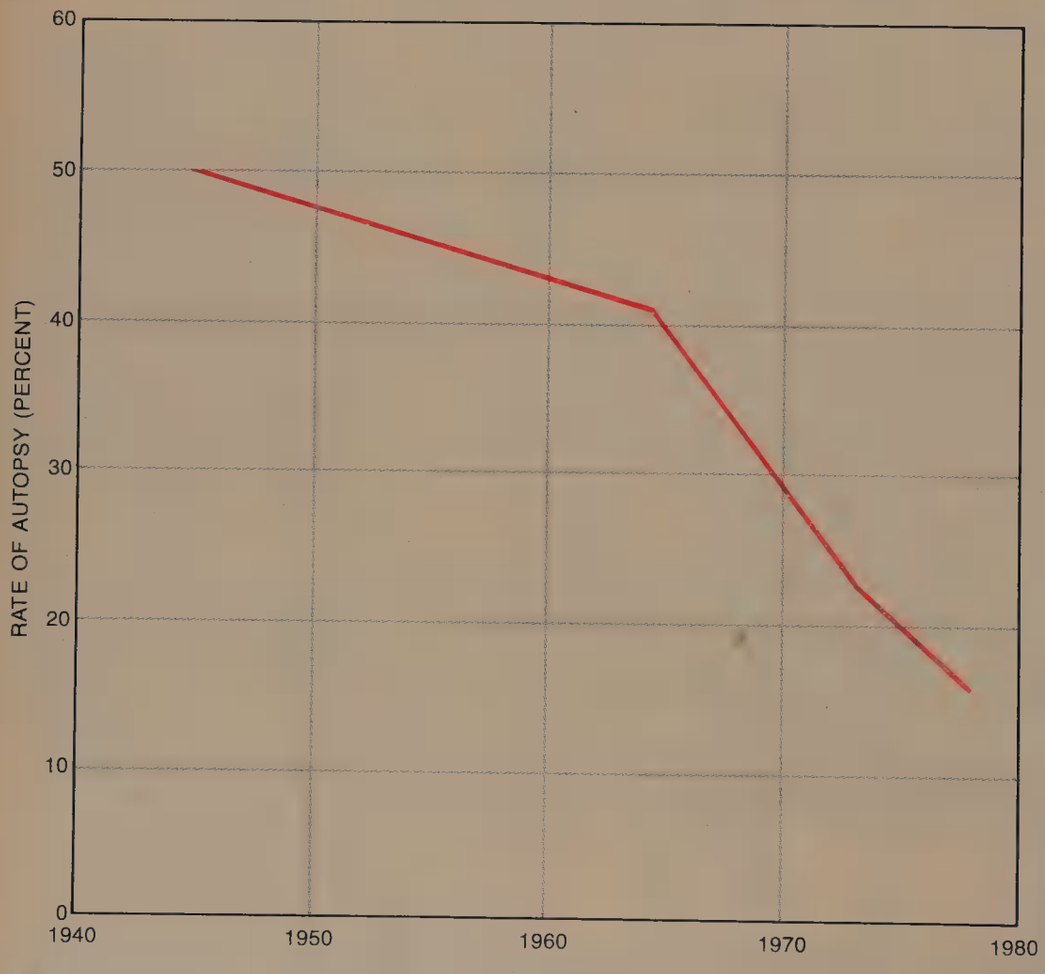
Senior pathologists must again de-



LEFT KIDNEY is shown (*left*) after it has been separately dissected from the organs associated with it. An incision has been made, as indicated by the broken line, to open the kidney for an examination of its interior. No disease is found. The descending tube is the ureter.



SKEWED PATTERN OF AUTOPSIES is evident in a comparison of the six categories for cause of death in which the autopsy rate was 50 percent or higher in 1977 with five categories that accounted for far more deaths and included diseases that are among the nation's major health problems. The figure at the end of a bar shows the number of deaths in that category.



DECLINING RATE OF AUTOPSY in the U.S. is charted as a percentage of the number of deaths. One reason for the decline is belief among physicians that little new is learned at autopsy, yet several diseases have recently been found or elucidated by the procedure.

velop and demonstrate skills in the performance of autopsies and must be able to solve clinical problems as well as morphologic ones. It used to be that physicians embarking on a career in pathology had at least one year of clinical experience. Similarly, clinicians spent some part of their training in a pathology department, often doing autopsies. This interchange of experience was thought to be particularly valuable for the physician in training and to be ultimately valuable for the patient. Pathologists were thought to gain a greater appreciation of the relevance of their activities, together with an ability to relate understandingly to their clinical colleagues. A step or two backward might be an advance.

An autopsy is a significant expense for a hospital, costing roughly $1,000, and it does not produce any income for either the hospital or the pathologist. The autopsy should be adequately funded so that it is not always second or lower in the line of priorities. Health-insurance agencies could benefit by providing direct reimbursement for the autopsy, since it yields many clues about disease and the causes of disease that can ultimately help to control the rising costs of health care.

At one time hospitals had to demonstrate an autopsy rate of 20 percent in order to be accredited. The requirement was eliminated in the 1960's. Data are lacking on what would be the most cost-effective and beneficial autopsy rate. Certainly it would not be worthwhile to do an autopsy in every case of death. William Roberts has suggested that some statistically determined sample of all people dying in a hospital should be autopsied rather than only those thought to present diagnostic problems, in order to achieve a reliable means of quality control. The ideal autopsy rate for a tertiary-care, teaching hospital is probably 100 percent. It is clearly not easy to determine the lowest acceptable rate.

John R. Carter, former director of the Institute of Pathology at Case Western Reserve University, is leading an effort by the College of American Pathologists to develop a computerized National Autopsy Data Bank, which would provide a central repository of pathological, biomedical, demographic and epidemiologic information. By this means the vast fund of medically and socially useful information yielded by autopsies would be readily accessible to everyone who might benefit from it. The data bank would surely bring about a far wider application of the inscription in the amphitheater where dissections were performed at the University of Bologna: *Hic locus est ubi mors gaudet succurere vitae* ("This place is where death rejoices to come to the aid of life").

THE INPUT/OUTPUT STRUCTURE OF THE UNITED STATES ECONOMY

WHAT MAKES THE U.S. ECONOMY TICK?

The editors of SCIENTIFIC AMERICAN have prepared a wall chart displaying for the 1980's the Input/Output Structure of the U.S. Economy based on the latest interindustry study from the U.S. Department of Commerce.

The SCIENTIFIC AMERICAN Input/Output wall chart does for economics what the table of elements does for chemistry. It answers at a glance questions about the linkage between the microeconomics of the firm and the macroeconomics of the system; about the web of technological interdependencies that tie industry to industry; about the industry-by-industry direct and indirect consequences of swings in public and private spending; about the impact of change in technology, and about any other topic you can think of. You are rewarded by surprise as well as by confirmation of your hunches. For teaching and practical and theoretical studies, here is a powerful, graphic tool.

In the familiar format of the SCIENTIFIC AMERICAN Input/Output wall charts for the 1960's and 1970's, the wall chart for the 1980's measures 65″ × 52″ and is printed in eight colors. Each of the nearly 10,000 cells in the 97-sector interindustry matrix shows (1) the interindustry commodity flow, (2) the direct input/output coefficient and (3) the "inverse" coefficient. Where the direct input/output coefficient exceeds .01, the cell is tinted in the color code of the industrial bloc from which the input comes. This device, combined with triangulation of the matrix, brings the structure of interindustry transactions into graphic visibility.

A supplementary table displays, industry by industry, the capital stock employed; the employment of managerial, technical-professional, white-collar and blue-collar personnel; the energy consumption by major categories of fuel, and environmental stress measured by tons of pollutants.

The editors of SCIENTIFIC AMERICAN are happy to acknowledge the collaboration, in the preparation of this wall chart, of Wassily Leontief, originator of input/output analysis—for which contribution to the intellectual apparatus of economics he received the 1973 Nobel prize—and director of the Institute for Economic Analysis at New York University.

Packaged with the chart is an index showing the BEA and SIC code industries aggregated in each of the 97 sectors.

SCIENTIFIC AMERICAN
415 Madison Avenue, Dept. I/O, New York, N.Y. 10017

YES, send me the new SCIENTIFIC AMERICAN 1980's INPUT/OUTPUT wall chart—printed in eight colors on heavy paper measuring 65″ × 52″.

I understand my chart(s) will be shipped prepaid, uncreased*, flat-rolled and in a sturdy mailing tube. (Allow 4-6 weeks for delivery)

I enclose a check for $____ for ____ chart(s) at $35 each, saving me $5 each in handling charges.

Bill me for $____ for ____ chart(s) at $40 each.

Name_____

Company_____

Street Address_____

City, State, Zip Code_____

*Charts shipped international require one fold. 3/83

Make check payable to SCIENTIFIC AMERICAN. Please add sales tax for California, Illinois, Massachusetts, Michigan and New York.

THE AMATEUR SCIENTIST

How to analyze a city traffic-light system from the outside looking in

by Jearl Walker

If you drive an automobile in a city, you are sure to encounter a time when you have to stop for a red light at almost every intersection. Your progress is slow, your time is wasted and you may find yourself in a massive traffic jam. You think if only the sequence of traffic signals along the route were synchronized, you might be able to go indefinitely without stopping.

Actually in most cities the important traffic routes now have synchronized light systems, particularly during rush hours. When the drivers travel at a certain speed, they are supposed to find a green light at each intersection. In heavy traffic this sequencing forces the cars into what traffic analysts call platoons. In the gaps between them the traffic on perpendicular streets crosses the main street or merges onto it. On a properly designed route the sequence of lights moves the traffic efficiently without unduly delaying the perpendicular traffic.

To examine a synchronized light system I studied the lights along Carnegie Street in Cleveland. To be sure, I could have found out much of what I wanted to know by consulting city officials, but it would not have been as much fun. Besides, there is always the possibility that an independent analysis can come up with some useful ideas. You may want to try the same kind of analysis on your own local traffic system.

The lights along Carnegie Street are controlled by a preset electronic system rather than by a feedback system sensitive to the density of the traffic. The street is a major artery between downtown Cleveland and the eastern suburbs and is heavily traveled in the morning and afternoon rush hours. In the morning rush hour the westbound traffic has four lanes and the eastbound has two; in the afternoon rush hour the traffic is one-way eastbound. For the rest of the day the street has three-lane traffic in each direction.

Carnegie Street runs through an area that includes the Cleveland Play House and a variety of stores, homes and warehouses. Side streets, each three lanes wide, connect Carnegie with parallel streets that also carry rush-hour traffic. Since much traffic crosses between the major streets, the red lights on the side streets cannot stay on too long or the queues on them may grow to the point where they block the major streets.

I studied the system of traffic lights on Carnegie Street in the early stages of the afternoon rush hour. Beginning at the intersection of Carnegie and East 71st streets, I measured the distance from one intersection to the next as I walked along Carnegie in the direction of the traffic flow. At each intersection I timed the duration of the red, yellow and green lights. In addition as I proceeded I recorded the interval between the start of the green light at one intersection and the start of the green light at the next.

My tool for ascertaining times was a digital watch with a stopwatch feature and a particularly helpful split-time readout. I started the watch when the yellow light came on at an intersection. When the red or green light came on, I pushed the button for the split time. This function briefly held the display so that I could record the time in my notebook. The stopwatch was still running, however, and after about five seconds the display again recorded the passing seconds as before.

After many observations I averaged the readings for lights of each color in order to reduce the error arising from variations in the length of time it took me to react to a change of the light and to push the appropriate buttons on the watch. I also averaged my measurements of the time between the beginning of the green light at one intersection and that at the next intersection.

To measure the distances between intersections I counted paces between the stop lines at successive intersections. Later at home I walked at approximately the same pace across my yard. After measuring the distance with a meterstick I was able to convert my measurements of distance along Carnegie into units of meters. I walked the width of my yard rather than taking a single step in order to reduce the error that might result from variations in the length of my stride. Even so, I figure that the error in my measurements of the distance between the intersections is between 10 and 20 percent. Although surveying would be far more accurate, the added precision in my results would not be worth the extra effort.

The measurements are summarized in the illustration on the opposite page. The intersection of Carnegie and East 71st is represented at the lower left. The flow of traffic along Carnegie is upward in the drawing of the street. The other intersections are laid out along the drawing.

The horizontal scale of the graph represents the time of the lights. For example, the green light at Carnegie and East 71st is plotted as beginning at a time of five seconds. This is the bench mark from which that light and the ones at succeeding intersections were measured. I found that the green light at East 71st stays on for an average of 46.4 seconds. Hence the graph shows the onset of the yellow light at 51.4 seconds, the sum of 46.4 seconds and the bench mark of five seconds. The graph indicates similarly the start of the red light and the times for the next cycle of green, yellow and red.

The green light at Carnegie and East 77th, the next intersecting street, begins 29.5 seconds after the green begins at East 71st. Therefore the graph shows the start of the green phase at East 77th to be at 34.5 seconds, the sum of 29.5 seconds and the bench mark of five seconds. Since the green light there lasts for 48.6 seconds, the beginning of the yellow light is at 83.1 seconds on the graph. The color cycles for the other intersections are similarly plotted with respect to the bench mark for the green light at East 71st.

During the afternoon rush hour platoons of cars form on Carnegie near East 71st or somewhat closer to the downtown section. Usually a platoon moves eastward at a nearly constant speed. A platoon leader could race from intersection to intersection, but it would be to little advantage: over an extended time and distance one cannot go faster than the sequence of green lights allows. If the leader leaves one intersection at maximum acceleration and excessive speed, the car reaches the next intersection much too early for the green light.

The graph indicates the motion of a platoon leader traveling at constant speed through the system of six intersections. The leader begins to move when the light at East 71st turns green and then travels along Carnegie at a constant speed to avoid red lights. According to my measurements, the light cycles are

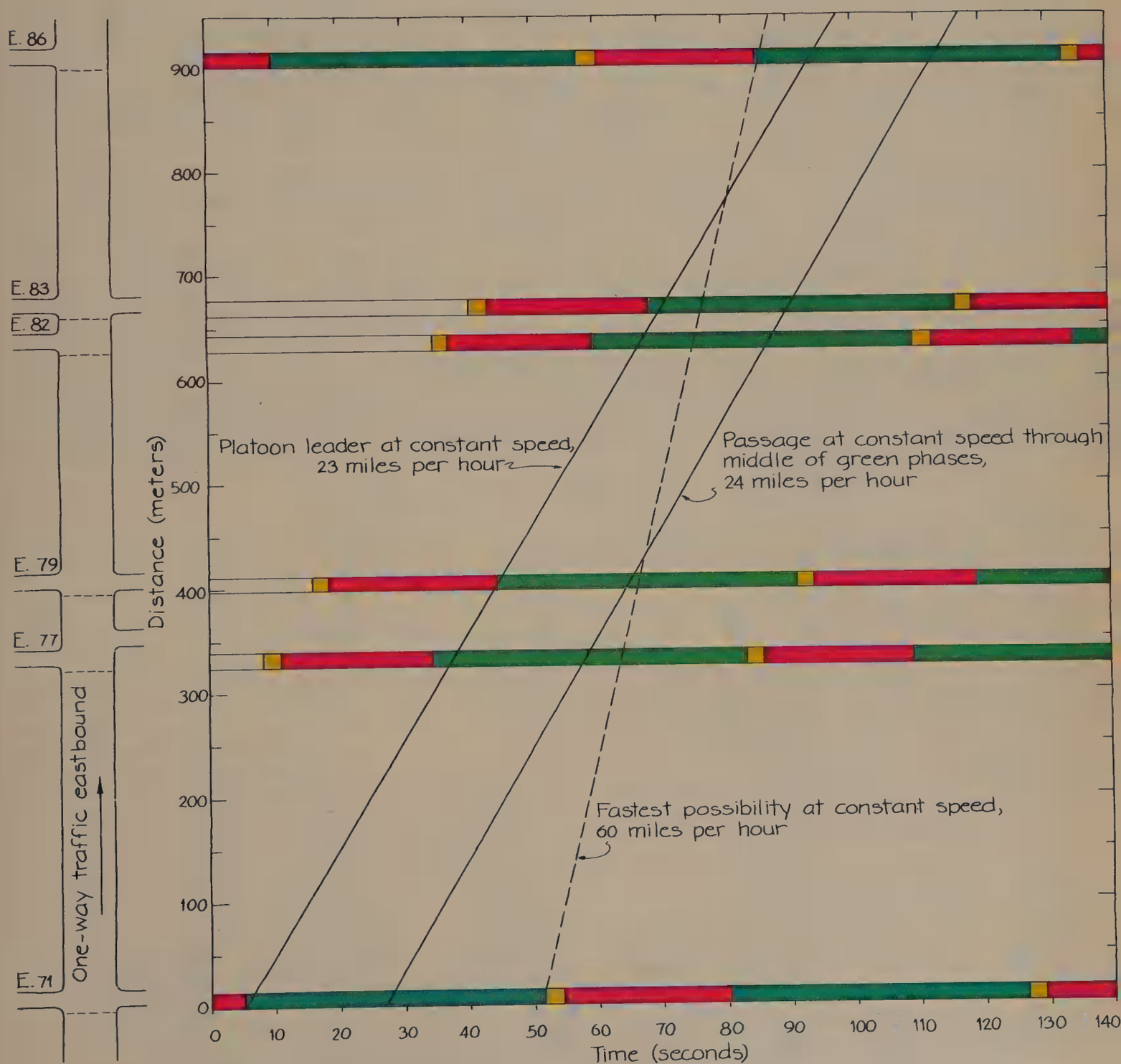
not synchronized perfectly for the leader. If the driver is to pass through the intersection at East 79th just at the onset of the green light, he reaches East 82nd about five seconds after the green light there has come on. Perhaps the mismatch is part of the design. It would allow the traffic merging onto Carnegie from East 79th to clear the intersections at East 82nd and East 83rd before the next platoon arrives.

The graph also indicates two other hypothetical drivers moving through the system at constant speed. One of them passes through each intersection at approximately the middle of the green phase. The other driver passes through East 71st precisely at the end of the green phase and then races to pass through East 86th just as the green light there comes on. The speed of the platoon leader and the other two drivers can be measured from the graph: the speed in meters per second is the slope of the line representing each driver.

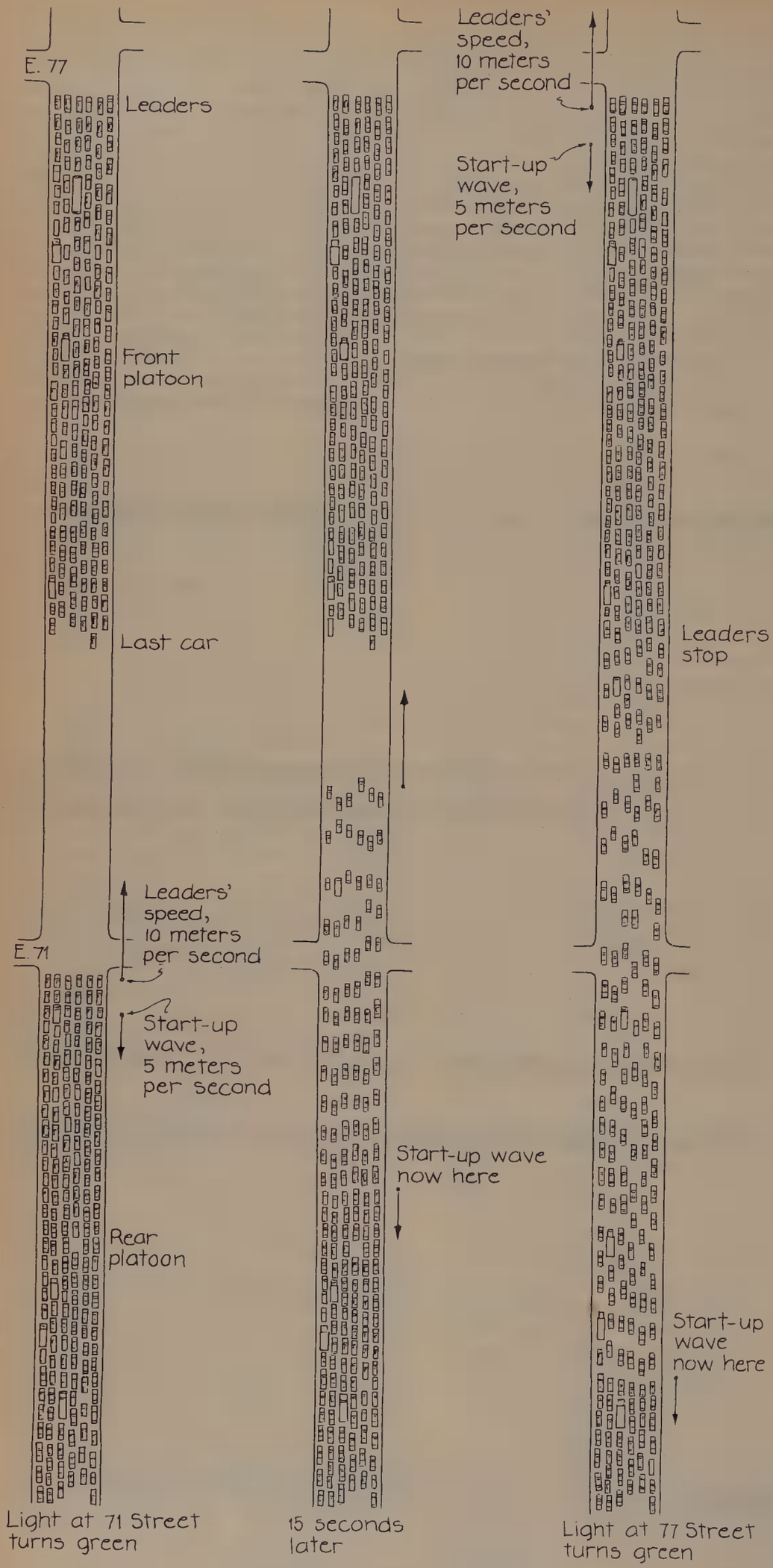
The platoon leader travels at 10.3 meters per second, which is about 23 miles per hour. The driver passing through each intersection at the middle of the green phase travels at 10.8 meters per second, which works out to approximately 24 m.p.h. The driver who races through the system in the least time moves at 26.6 meters per second, which is about 60 m.p.h. (The speed limit on Carnegie Street is 25 m.p.h.)

Compared with what happens in real traffic my law-abiding drivers are somewhat slow. When I drive through the rush-hour system on my way home every day, my platoon usually moves at about 28 m.p.h. Three factors contribute to the discrepancy between this speed and the speed computed from the graph. If the traffic flow is to be steady, a platoon leader must have a green light one or two seconds before arriving at an intersection, otherwise caution makes him slow down. The graph also ignores the time it takes the platoon leader to respond to the onset of the green light at East 71st. I figure that the reaction time is about a second.

The third factor is that the graph over-



The system of traffic signals on Carnegie Street in Cleveland during the afternoon rush hour



Conditions that can lead to a traffic jam

looks the time required to accelerate a car from rest to the constant speed that takes the platoon through the system. When I am the platoon leader, I need about three seconds to reach cruising speed. These three factors effectively reduce the time a platoon leader has to travel between intersections. Therefore he can drive at a constant speed of somewhat more than 23 m.p.h. without having to stop at any intersection.

At times I have driven through the system after Carnegie has become one-way but before the volume of traffic has built up. If I pass through East 71st near the end of the green phase, I can go faster than 50 m.p.h. through several of the intersections but must stop for the red light at the intersection past East 86th. Almost every time I have exceeded the speed limit I have been passed by another driver. Hence it is possible to travel through the system at a constant speed of about 60 m.p.h.

During a rush hour platoons of cars move along Carnegie about every 75 seconds, which is the interval between successive green lights at East 71st. Although the lights along the route are not synchronized perfectly, the system works well enough so that the platoons are not too large for the distance between successive intersections. In fact, the green phase for most of the intersections could be decreased by 10 or 15 seconds without delaying the average platoon. The green phases are apparently kept as long as they are in case something goes wrong. The extra time is needed if an accident clogs one of the lanes of traffic or the rush-hour traffic is notably heavy.

Most of the traffic on Carnegie in the early stage of the afternoon rush hour tends to stay toward the right side of the street, although many drivers avoid the rightmost lane because of the possibility of coming on a car that is disabled or illegally parked. As the volume of traffic builds up, the lanes become almost uniformly filled.

Until that time a platoon can normally maneuver around an obstacle in one of the lanes. When a lane of traffic slows, drivers at the rear of the queue pull into other lanes. An obstacle tends to force the traffic into a more uniform distribution among the lanes. Trouble develops when the other lanes are already full. Then the queue that forms behind an obstacle can disrupt the scheme designed to move platoons of cars through the system of lights.

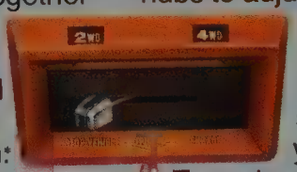
The plan of the light system on Carnegie seems to assume that when a platoon leader approaches an intersection, the preceding platoon of cars will have moved on. If part of that platoon still blocks the way, the leader of the new one must slow down or stop. The entire timetable of lights can be thrown off.

Why drive an ordinary wagon when you can drive the Ultimate Wagon?

Any similarity between the Jeep Wagoneer and a conventional two-wheel drive station wagon is purely coincidental.

Because no conventional full-size wagon gives you the

Wagoneer's four-wheel drive traction and security, together with Wagoneer's two-wheel drive economy... better EPA estimated MPG than any full-size two-wheel drive wagon*. Add to that the comfort and luxury you'd expect to find only in the plushest automobiles, and it's hard to call Wagoneer anything but *beautiful*.



Introducing Selec-Trac.

Now Wagoneer offers even more convenience: two-wheel drive or full-time four-wheel drive at your fingertips. No wheel hubs to adjust. And no need to leave the driver's seat. Just flick the dash-mounted Selec-Trac switch, and you're on your way.

Two-wheel drive for improved fuel economy. Four-wheel drive for sure traction in rain, sleet or snow, either on-road or off-road...even at faster highway speeds.

So why drive a conventional wagon when you can drive the Ultimate Wagon. Jeep Wagoneer Limited.

18 EST* EPA MPG 23 HWY EST

Jeep® Wagoneer Limited. The Ultimate Wagon. SAFETY IS A SNAP

AT AMERICAN MOTORS

*Use these figures for comparison. Your results may differ due to driving speed, weather conditions and trip length. Actual highway mileage lower.

Jeep Corporation, ■ subsidiary of American Motors Corporation.

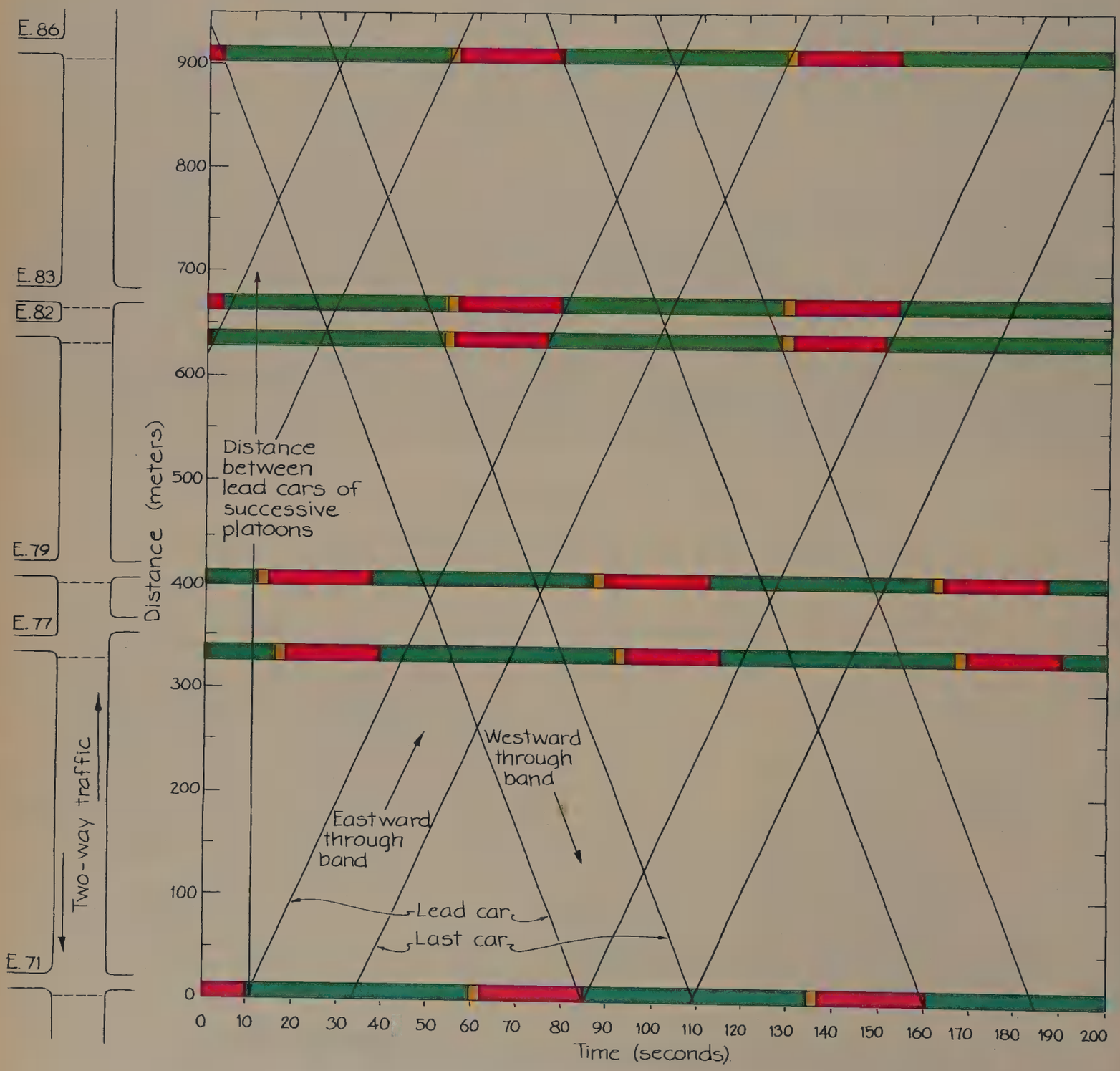
The delay may mean that part of the new platoon will not pass through the intersection before the next one arrives. If this happens, the problem becomes progressively worse.

The illustration on page 140 shows two platoons on Carnegie Street. The leading platoon is stopped at a red light at East 77th. The rear platoon has just got a green light at East 71st. First the leaders of the rear platoon begin to move into the intersection at East 71st. Then the cars behind the leaders move. This start-up motion travels the length of the platoon as a wave. Finally the last cars of the rear platoon move forward.

Once the leading platoon gets a green light it too will have a start-up wave propagating to its last cars. If they begin to move before the leaders of the rear platoon arrive, the light system works fine and all the cars move through it without unnecessary delay. Suppose, however, the rear platoon arrives too soon. Since its members must stop, the queue of cars gets longer and may eventually block the intersection at East 71st during the next change of lights there. The situation then has the makings of a traffic jam.

I have been caught in several traffic jams on Carnegie. Once it took me almost two hours to get through the six intersections from East 71st to East 86th. A heavy snowfall had delayed the early platoons of cars traveling through the system. Parts of the platoons were halted by red lights. Because of the slow driving on the snowy streets, the queues at the intersections lengthened as platoons from the rear arrived before the preceding platoons had moved on. The density of cars quickly became maximum: cars were bumper to bumper for miles along Carnegie. Since the intersections were blocked, the traffic on the perpendicular streets also became jammed. If traffic flow is considered to be an example of hydrodynamics, the traffic on that miserable winter afternoon was frozen solid.

For the situation depicted in the illustration on page 140 what length of the leading platoon causes the rear one to



The timing of signals for two-way traffic at midday

be delayed? The critical length can be determined in terms of the distance between the intersections; it depends on the rate at which the start-up wave travels the length of the leading platoon. If the wave travels much slower than the speed at which the rear platoon approaches, even a short leading platoon can cause trouble.

To measure the speed of the start-up wave I stationed myself at intersections along Carnegie during a period of heavy traffic. I chose intersections that normally have queues of 10 cars or more. When the light for a queue turned green, I started my stopwatch. I stopped it when the last car in the queue began to move. I estimated the total distance between the front of the first car and the rear of the last. Dividing that distance by the time elapsed on the stopwatch, I estimated the speed of the wave. After many observations I found that the average speed was about five meters per second, or about half the designed cruise speed.

This finding means that the critical length for the leading platoon is only a small fraction of the total distance between East 71st and East 77th. If that platoon is more than about three car lengths long, the leaders of the rear platoon will be delayed. If both platoons are much longer than that, the rear platoon will not clear the intersection at East 71st before the next red light there.

Traffic jams along Carnegie might be avoided if the sequence of lights could be altered when the queues become too long for the normal rush-hour sequencing. The longer the leading platoon at East 77th is, the less head start the rear platoon should be given. If the length of the leading platoon is about a third of the distance between the two intersections, the green lights should come on simultaneously. If the leading platoon is even longer, the light at East 77th should come on before the one at East 71st, an order that is the reverse of the normal rush-hour sequencing. The head start given the leading platoon then allows the start-up wave to reach the last cars before the leaders of the rear platoon arrive.

In the bottom illustration on page 145 is an equation bearing on the platoon length at which the normal sequence must be changed. The length of the platoon (x) is expressed as a fraction of the distance between intersections (d). It depends on the speed of the start-up wave (v_1) and the normal cruise speed (v) for the system. When the platoons are shorter than the switchover value, the rear platoon should get the head start. When they are longer, the leading platoon should be given the head start. If the platoons are as long as the switchover value, they should get simultaneous green lights.

Some systems of traffic lights can be

Which of these languages would you like to speak?

Mark the one you want to speak in 1 or 3 months' time

Afrikaans	☐	French	☐	Japanese	☐
American English	☐	German	☐	Norwegian	☐
Arabic	☐	Greek (Modern)	☐	Polish	☐
Chinese	☐	Hebrew (Modern)	☐	Portuguese	☐
Czech	☐	Hindi	☐	Russian	☐
Danish	☐	Indonesian	☐	Spanish	☐
Dutch	☐	Irish	☐	Swedish	☐
Finnish	☐	Italian	☐	Welsh	☐

A Linguaphone Course makes you feel at home in almost any country. You'll never miss the meaning of conversations or be at a loss for words.

- It must work—over 4 million Linguaphone students in 88 countries speak a second language FLUENTLY.
- Proven learning success. You LISTEN to real conversations on cassettes...UNDERSTAND what you hear by following illustrated textbooks. HOLD CONVERSATIONS with the speakers.
- **You start speaking the very first lesson.** You learn at your own convenience.
- It's like having a private tutor.
- You get a complete, professional language program at little cost.
- You gain a good, working vocabulary.
- In just 2 to 3 months you can speak another language with complete confidence.
- You develop an authentic accent. Only native-born speakers are used.

Linguaphone — The Language Masters

MONEY BACK GUARANTEE—28 DAY FREE TRIAL

World Language Courses, Inc. Dept. 250 313 Nolana Ave. McAllen, Texas 78501

FREE INFORMATION: Please mail me FREE information about learning the languages I have checked. FREE brochure.

Name (please print) _____

Street _____

City _____ **State** _____ **Zip** _____

ANTIQUITY

A Periodical Review of Archaeology
edited by Glyn Daniel

Founded in 1927 by O. G. S. Crawford, ANTIQUITY has appeared regularly ever since and won acclaim the world over as the most authoritative journal in its field. While written by specialists, the articles, notes and reviews are popular in character and indispensable to all interested in the development of man and his past.

Professor Glyn Daniel, Faculty of Archaeology and Anthropology in the University of Cambridge, and Fellow of St John's College, has been Editor since 1956.

The annual subscription, postage included, is $35. Subscription forms and bankers' orders are available on request from

ANTIQUITY PUBLICATIONS LIMITED
Heffers Printers Ltd, King's Hedges Road,
Cambridge, England CB4 2PQ

Just Published
1983 Edition of
PHYSICIANS' DESK REFERENCE

The definitive reference on prescription drugs. Used by over 400,000 physicians throughout the U.S. Contains ALL information required to be available to prescribers according to FDA guidelines. Each drug listing includes descriptions of chemical makeup, actions, contraindications, adverse reactions, dosage forms, action in the body and more. Nearly 2000 packed pages give you the most complete, in-depth, authoritative source of prescription drug information available to the professional. This is not a watered down superficial treatment. The reader can skim an entry or probe deeply into the specifics of the product.

Highest quality full color photographs, actual size, show many of the products listed in the book. An invaluable guide to drug identification, this section has proven invaluable in emergency situations.

Four indices allow the reader to find a drug by its manufacturers name, its product name, the category of ailment for which it is primarily intended or its generic or chemical name. These, along with the color photographic section, allow the reader to locate a product regardless of the amount of information available at the start of a search.

DETACH ALONG DOTTED LINE AND MAIL TODAY

1. ☐ For greatest savings, include your check for $19.95 per copy.
2. ☐ I prefer to be billed later at a cost of $21.95 per copy.

NAME

ADDRESS

CITY

STATE ZIP

SEND YOUR ORDER TO:
Physicians' Desk Reference
P.O. Box 58-SA
Oradell, New Jersey 07649

SA-383

changed in sequence when the queues become too long. The change could be preprogrammed into the system if long queues can be expected. In some systems the light sequence is affected by devices laid in the roadway to sense traffic flow. If Carnegie had such a system, my fellow prisoners and I might have been spared the ordeal of traveling only five blocks in two hours.

In the late morning and early afternoon Carnegie has three lanes of traffic in both directions. The sequencing of the lights should be different from what it is in the afternoon rush hour so that the westbound traffic will not be stopped too many times. I repeated my measurements of the light cycles at about noon. The results are given in the illustration on page 142, which shows the green light at East 71st starting at a time of 10 seconds. Thus 10 seconds is the bench mark for the color cycles of the intersections.

At each intersection the duration of the three colors of light is the same as it is during the rush hour, but the time between the onset of green lights at successive intersections is not the same. For example, during the rush hour the green light at East 79th comes on 15.4 seconds before the green light at East 82nd. The delay allows a platoon enough time to reach East 82nd before the green appears. For two-way traffic the system must be synchronized to minimize the delay for both directions of flow without forcing the perpendicular traffic to wait too long. At noon the green at East 82nd and East 83rd is already on when the green starts at East 79th. This system allows the westward traffic to travel uninterrupted through the intersections.

The illustration also shows "through bands," which are often put in such a graph of distance v. time. Here a through band indicates how a platoon traveling at a constant speed can pass through a system of intersections without pause. Although in light traffic cars could proceed at various speeds, the through bands are instructive to someone deciding how to synchronize the lights for most of the traffic. The illustration has two sets of through bands, one for eastward flow (the bands running upward to the right) and another for westward flow (the bands running downward to the right).

A through band consists of two parallel lines drawn as far apart as the graph for a traffic route allows. One of the eastward through bands in the illustration begins with the onset of the green light at East 71st. A straight line passes through the green phases of the other intersections up through East 86th. (The line is straight because it represents traffic moving at a constant speed.) Since the traffic is to flow along the route as quickly as it can without breaking the speed limit, the line on the graph should be as steep as possible. The line also must avoid any of the red phases because it represents uninterrupted traffic. Hence it skims across the end of the red phases at East 82nd and East 83rd.

The right side of the through band is parallel to the left side because it represents the rear cars of the platoon, all of whose members are traveling at the same speed. This line also is drawn to avoid any of the red phases. In the illustration the right side of the eastward through band skims the onset of red at East 86th.

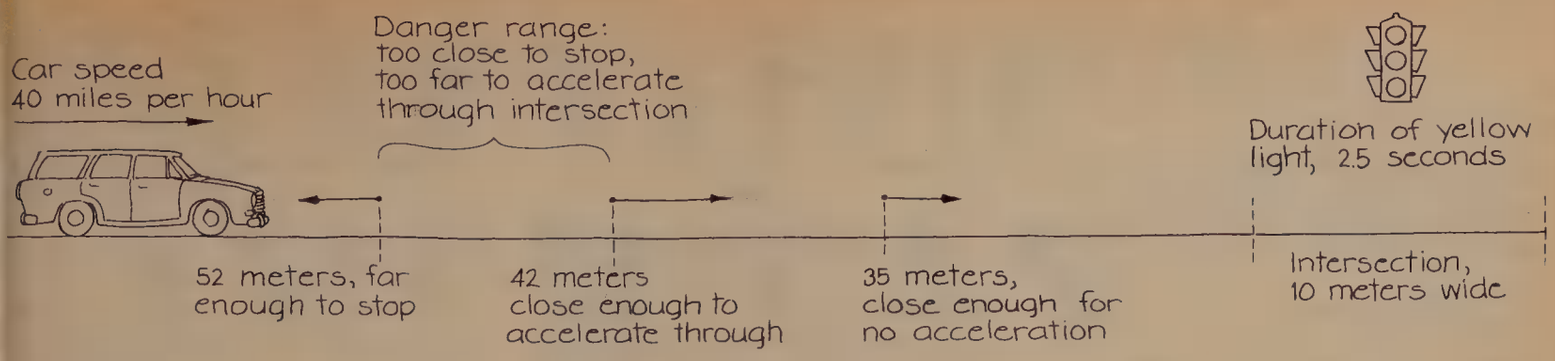
Suppose the eastward traffic along Carnegie at noon consists only of platoons of cars. The cruise speed of a platoon can be computed from the slope of the lines of the eastward through bands. If the traffic is to be uninterrupted, the platoon must begin at the onset of the green at East 71st and travel through the system of lights at a speed of 9.6 meters per second (equivalent to 21.5 m.p.h.).

How much later the rear of the platoon must pass through the intersection at East 71st is indicated by the right side of the through band. Apparently the rear must be no later than about 26 seconds after the green begins at that intersection. If a car in the platoon passes through the intersection later in the green phase there, it will be stopped by a red light at East 86th if not sooner. The through band during a rush hour is much wider, encompassing the entire green phase at East 71st. Thus when the Carnegie light system changes to the sequence for rush-hour traffic, the platoon speeds are not altered but the platoons can be much longer.

The distance between adjacent platoons can also be determined from the illustration. Start at the left side of one eastward through band, say at the bench mark for the onset of the green at East 71st. From the vertical axis read the distance to the left side of the through band between East 83rd and East 86th for that same time. The reading (about 700 meters) is the distance between the leading cars of the successive platoons. The distance between the last cars of the leading platoon and the first cars of the platoon beginning at East 71st is about 500 meters.

The illustration also includes information about the westward through bands. The left side of a band represents the travel of the first cars of a platoon. The right side represents the travel of the platoon's last cars. The westward and eastward through bands are approximately the same size, indicating that the system of lights is designed to pass traffic equally well in those directions at this time of day.

What is a suitable duration for the yellow light? It must be short enough to not



Strategies for a driver approaching a yellow light

delay the traffic unduly and long enough to allow a driver either to stop properly or to pass through the intersection before the red light comes on. The appropriate duration depends partly on the speed limit of the street, which determines a driver's ability to avoid being in the intersection when the red comes on. The higher the speed limit is, the longer the yellow light should be. This relation is sometimes ignored in setting the phases of a traffic light.

Assume that a yellow light lasts for 2.5 seconds, which is about the shortest duration I have found in Cleveland. Assume further that when the yellow first comes on, the driver immediately responds by braking to a stop. The minimum distance required for the stop depends on two factors: the car's initial speed (v_0) and the acceleration (a) provided by braking. (A physicist refers to any change in velocity as an acceleration; in common parlance the slowing of a car is called deceleration.) The stopping distance is computed as being the speed squared divided by twice the acceleration.

The acceleration of a braking car can vary greatly because it depends on the weather, the type of pavement, the condition of the tires and brakes and the skill of the driver. Usually the acceleration is initially between .9 meter and 3.1 meters per second per second. In the last stage of braking the acceleration is about 3.7 meters per second per second. In an emergency the acceleration could be as high as 6.1 meters per second per second, but braking at that rate is uncomfortable and unnerving.

Assume an acceleration of 3.1 meters per second per second. Suppose the driver is traveling at 40 m.p.h., which is 17.9 meters per second. The minimum stopping distance is 52 meters. If the driver is to stop before entering the intersection, the car must be no closer to the intersection than that distance. If it is closer, the driver has no reasonable choice except to continue through the intersection with no attempt at braking. Perhaps he should even accelerate in order to avoid being in the intersection when the red light comes on.

How successful can the driver be in such an acceleration? What he can do depends not only on the car's pickup but also on the initial distance of the car from the intersection. If the intersection is too far away, the driver cannot reach the far side before the red light comes on. The formula in the bottom illustration on this page yields the maximum distance the car must be from the intersection if the attempt to get through is to be successful. The formula encompasses the duration of the yellow light (t) and the acceleration (a) and initial speed (v_0) of the car. It also requires the width of the intersection (s) because the driver must reach the far side while the yellow light is still on.

A typical car can accelerate at from one meter to 2.2 meters per second per second. A sports car can probably accelerate at twice that rate. Consider a car with an acceleration of 2.2 meters per second per second. Assume that the driver responds instantly to the onset of the yellow light. Take the duration of the yellow light as being 2.5 seconds and the width of the intersection as being 10 meters. If the car is initially traveling at 40 m.p.h., the maximum distance between it and the intersection should be about 42 meters; if it is more than that, the car will be in the intersection during part of the red light.

The driver may decide to go through the yellow light without accelerating beyond the speed limit. If the initial speed of 40 m.p.h. is maintained, the car should be no farther than 35 meters from the intersection.

I have chosen the above numbers in order to make a point. If the driver is to stop successfully, the car must be no closer than 52 meters from the intersection. If he is to race through at maximum acceleration, the car must be no farther than 42 meters from the intersection. Between those limits what should he do? In principle neither strategy will succeed. The dangerous region between the two limits is larger with higher initial speeds for the car. Under the same assumptions as before except for an initial speed of 55 m.p.h., the danger region is 39 meters long. If the driver decides not to accelerate, the danger region is 46 meters long.

I once drove through such an intersection on a highway that had a speed limit of 55 m.p.h. I found myself facing a yellow light with neither the space to stop nor the acceleration to race through before the red light came on. I was saved from the possibility of a collision only by a delay in the light system: the green light for the perpendicular traffic came on about a second or so after the yellow light ended.

This month's bibliography [*next page*] cites two articles in which teachers of physics have analyzed the dilemma of a short yellow light. You might want to examine yellow lights in your neighborhood. Please keep in mind that the values I have used for the accelerations of a car are assumed. The braking acceleration is particularly open to question. It depends on the coefficient of friction between the tires and the street. If the street is covered with ice, snow, rain or anything else that reduces friction, the coefficient can be much lower than normal. The distance required for stopping the car is then much longer. Therefore the danger region is also longer and the possibility of a collision between a car and a vehicle in the perpendicular traffic is greater.

Switchover length for successive platoons:

$$\frac{x}{d} = \frac{V_1}{(V_1 + V)}$$

Minimum distance to brake a car:

$$d = \frac{V_0^2}{2a}$$

Maximum distance to race through a yellow light:

$$d = v_0 t + \frac{1}{2} a t^2 - s$$

Maximum distance to drive at constant speed through a yellow light:

$$d = v_0 t - s$$

Equations for traffic flow

BIBLIOGRAPHY

Readers interested in further explanation of the subjects covered by the articles in this issue may find the following lists of publications helpful.

METAMAGICAL THEMAS

A NEW UCI LISP MANUAL. Edited by John R. Meehan. Lawrence Erlbaum Associates, Inc., 1979.

A MICRO-MANUAL FOR LISP—NOT THE WHOLE TRUTH. John McCarthy in *History of Programming Languages,* edited by Richard L. Wexelblat. ACM Monograph Series, Academic Press, 1980.

THE LAW OF THE SEA

OCEAN YEARBOOK. Edited by Elisabeth Mann Borgese and Norton Ginzburg. International Ocean Institute, University of Chicago Press, 1979–82.

OCEAN MANAGEMENT: A NEW PERSPECTIVE. John M. Armstrong and Peter C. Ryner. Ann Arbor Science Publishers/The Butterworth Group, 1981.

MICROPROGRAMMING

AN OVERVIEW OF FIRMWARE ENGINEERING. Scott Davidson and Bruce Shriver in *Computer,* Vol. 11, No. 5, pages 21–33; May, 1978.

MICROPROGRAMMED CONTROL AND RELIABLE DESIGN OF SMALL COMPUTERS. George D. Kraft and Wing N. Toy. Prentice-Hall, Inc., 1981.

MORE HARDWARE MEANS LESS SOFTWARE. Robert Bernhard in *IEEE Spectrum,* Vol. 15, No. 12, pages 30–37; December, 1981.

THE SCRIPT OF THE INDUS VALLEY CIVILIZATION

THE SCRIPT OF HARAPPA AND MOHENJODARO AND ITS CONNECTION WITH OTHER SCRIPTS. G. R. Hunter. Kegan Paul, Trench, Trubner & Co., Ltd., 1934.

A DRAVIDIAN ETYMOLOGICAL DICTIONARY. T. Burrow and M. B. Emeneau. Oxford University Press, 1961.

THE INDUS CIVILIZATION. Sir Mortimer Wheeler. Cambridge University Press, 1968.

THE ROOTS OF ANCIENT INDIA. Walter A. Fairservis, Jr. The University of Chicago Press, 1975.

THE INDUS SCRIPTS—TEXTS, CONCORDANCE AND TABLES. I. Mahadevan. Memoirs of the Archaeological Survey of India No. 77, Manager of Publications, Government of India, 1977.

A SKETCH OF COMPARATIVE DRAVIDIAN MORPHOLOGY. Kamil V. Zvelebil. Mouton Publishers, The Hague, 1978.

MITOCHONDRIAL DNA

SEQUENCE OF INTRONS AND FLANKING EXONS IN WILD-TYPE AND *BOX 3* MUTANTS OF CYTOCHROME B REVEALS AN INTERLACED SPLICING PROTEIN CODED BY AN INTRON. Jaga Lazowska, Claude Jacq and Piotr P. Slonimski in *Cell,* Vol. 22, No. 2, Part 2, pages 333–348; November, 1980.

MITOCHONDRIAL GENES. Edited by P. P. Slonimski, P. Borst and G. Attardi. Cold Spring Harbor Monographs, Vol. 12, 1981.

SEQUENCE AND ORGANIZATION OF HUMAN MITOCHONDRIAL GENOME. S. Anderson, A. T. Bankier, B. G. Barrell, M. H. L. de Bruijn, A. R. Coulson, J. Drouin, I. C. Eperon, D. P. Nierlich, B. A. Roe, F. Sanger, P. H. Schreier, A. J. H. Smith, R. Staden and I. G. Young in *Nature,* Vol. 290, No. 5806, pages 457–465; April 9, 1981.

THE FUTURE OF THE UNIVERSE

THE FIRST THREE MINUTES: A MODERN VIEW OF THE ORIGIN OF THE UNIVERSE. Steven Weinberg. Basic Books, Inc., 1977.

TIME WITHOUT END: PHYSICS AND BIOLOGY IN AN OPEN UNIVERSE. Freeman J. Dyson in *Reviews of Modern Physics,* Vol. 51, No. 3, pages 447–460; July, 1979.

EFFECTS OF PROTON DECAY ON THE COSMOLOGICAL FUTURE. Duane A. Dicus, John R. Letaw, Doris C. Teplitz and Vigdor L. Teplitz in *The Astrophysical Journal,* Vol. 252, No. 1, pages 1–9; January 1, 1982.

MEMORY IN FOOD-HOARDING BIRDS

FOOD STORING BY MARSH TITS. R. J. Cowie, J. R. Krebs and D. F. Sherry in *Animal Behaviour,* Vol. 29, pages 1252–1259; 1981.

MEMORY FOR THE LOCATION OF STORED FOOD IN MARSH TITS. D. F. Sherry, J. R. Krebs and R. J. Cowie in *Animal Behaviour,* Vol. 29, pages 1260–1266; 1981.

AN EXPERIMENTAL ANALYSIS OF CACHE RECOVERY IN CLARK'S NUTCRACKER. S. B. Vander Wall in *Animal Behaviour,* Vol. 30, pages 84–94; 1982.

OSCILLATING CHEMICAL REACTIONS

OSCILLATIONS IN CHEMICAL SYSTEMS, II: THOROUGH ANALYSIS OF TEMPORAL OSCILLATION IN THE BROMATE-CERIUM-MALONIC ACID SYSTEM. Richard J. Field, Endre Körös and Richard M. Noyes in *Journal of the American Chemical Society,* Vol. 94, No. 25, pages 8649–8664; December 13, 1972.

BIOLOGICAL AND BIOCHEMICAL OSCILLATORS: PROCEEDINGS OF THE CONFERENCE ON BIOLOGICAL AND BIOCHEMICAL OSCILLATORS, PRAGUE, 1968. Edited by B. Chance, E. K. Pye, A. K. Ghosh and B. Hess. Academic Press, 1973.

CHEMICAL OSCILLATIONS. G. Nicolis and J. Portnow in *Chemical Reviews,* Vol. 73, No. 4, pages 365–384; August, 1973.

A SYSTEMATICALLY DESIGNED HOMOGENEOUS OSCILLATING REACTION: THE ARSENITE-IODATE-CHLORITE SYSTEM. Patrick De Kepper, Irving R. Epstein and Kenneth Kustin in *Journal of the American Chemical Society,* Vol. 103, No. 8, pages 2133–2134; April 22, 1981.

MINIMAL BROMATE OSCILLATOR: BROMATE-BROMIDE-CATALYST. MiklósOrbán, Patrick De Kepper and Irving R. Epstein in *Journal of the American Chemical Society,* Vol. 104, No. 9, pages 2657–2658; May 5, 1982.

AUTOPSY

A HISTORY OF THE AUTOPSY. Lester S. King and Marjorie C. Meehan in *American Journal of Pathology,* Vol. 73, No. 2, pages 514–544; November, 1973.

THE AUTOPSY: ITS DECLINE AND A SUGGESTION FOR ITS REVIVAL. William C. Roberts in *The New England Journal of Medicine,* Vol. 299, No. 7, pages 332–338; August 17, 1978.

TO SEE OR NOT TO SEE: THE STATUS OF THE AUTOPSY AT THE MOUNT SINAI MEDICAL CENTER. Stephen A. Geller in *The Mount Sinai Journal of Medicine,* Vol. 46, No. 1, pages 33–38; January–February, 1979.

PROPOSAL FOR A NATIONAL AUTOPSY DATA BANK: A POTENTIAL CONTRIBUTION OF PATHOLOGISTS TO THE HEALTH CARE OF THE NATION. John R. Carter, Nancy P. Nash, Ronald L. Cechner and Ron D. Platt in *American Journal of Clinical Pathology,* Vol. 74, No. 4 Supplement, pages 597–617; October, 1981.

THE AMATEUR SCIENTIST

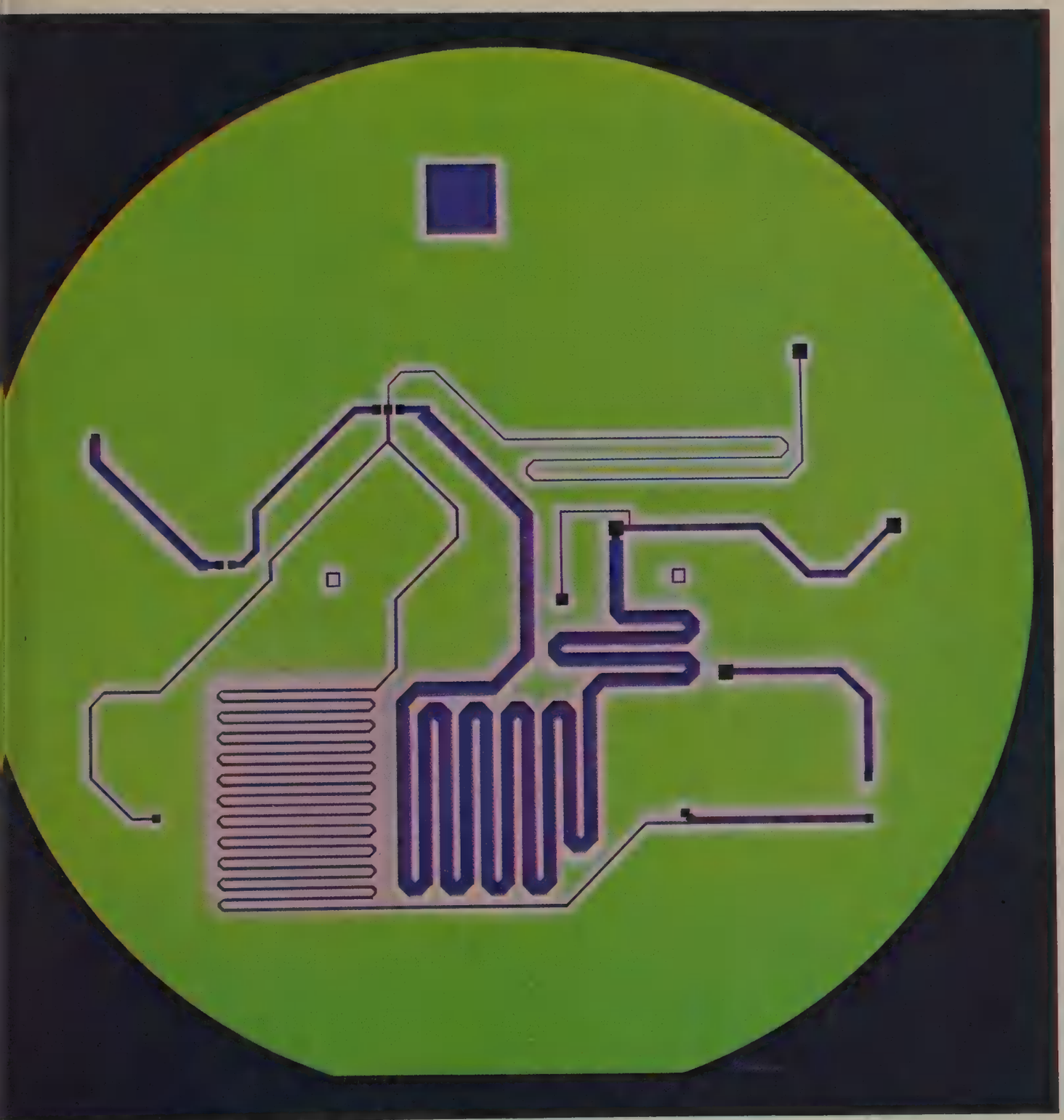
THE STOP-LIGHT DILEMMA. Howard S. Seifert in *American Journal of Physics,* Vol. 30, pages 216–218; 1962.

TO STOP OR NOT TO STOP—KINEMATICS AND THE YELLOW LIGHT. J. Fred Watts in *The Physics Teacher,* Vol. 19, No. 2, pages 114–115; February, 1981.

APR 4 '83

SCIENTIFIC AMERICAN

Marygrove College Library
Detroit, Michigan 48221
PLEASE DO NOT REMOVE



MICROMECHANICAL DEVICES

0 74820 08715 04

$2.50

April 1983

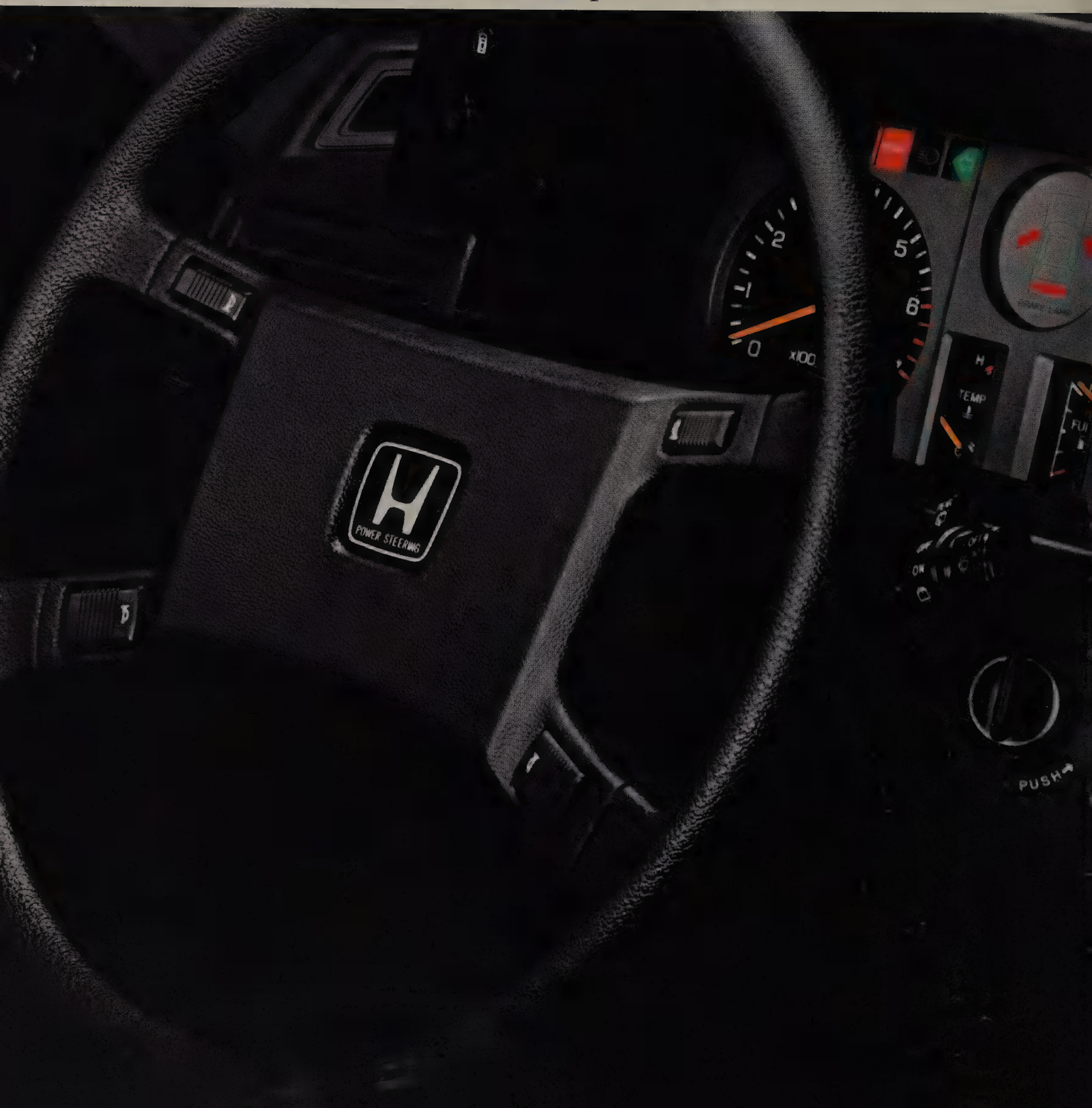
The simple luxury of quiet

In a noisy world, quiet is a luxury. One that comes standard with the Honda Accord LX Hatchback.

Wind and highway noise are reduced to a whisper by aerodynamic styling, triple door seals and a plush new interior.

Handling is remarkably smooth, the result of more Accord LX standard features. Like variable-assist power steering. Michelin steel-belted radial tires. Front-wheel drive. And 4-wheel independent suspension.

Even air conditioning is standard. Simply push a button and cool air flow through Honda's full-width instrumen panel vents.



Established 1845 **SCIENTIFIC AMERICAN** April 1983 Volume 248 Number 4

ARTICLES

37 **A NUCLEAR-WEAPON-FREE ZONE IN EUROPE, by Barry M. Blechman and Mark R. Moore** It could reduce the chance that battlefield uncertainties could lead to doomsday.

44 **SILICON MICROMECHANICAL DEVICES, by James B. Angell, Stephen C. Terry and Phillip W. Barth** Small machines can be built much as microelectronic circuits are.

56 **THE STRUCTURE OF QUARKS AND LEPTONS, by Haim Harari**
They have been considered elementary particles, but they may consist of still smaller entities.

78 **HOT SPRINGS ON THE OCEAN FLOOR, by John M. Edmond and Karen Von Damm**
The waters vented at sea-floor spreading centers deposit metal ores and sustain life without light.

94 **A WINDOW ON THE SLEEPING BRAIN, by Adrian R. Morrison**
REM (rapid eye movement) sleep is opened to study by eliminating its associated paralysis.

104 **EARLY FARMERS OF THE NORTH EUROPEAN PLAIN, by Peter Bogucki and Ryszard Grygiel** Excavations illuminate the transition from the Paleolithic to the Neolithic.

114 **CHEMICAL SIGNALS OF SOCIAL AMOEBAE, by John Tyler Bonner**
Two amoebae of the soil maintain their identity by emitting and responding to different chemicals.

122 **INTUITIVE PHYSICS, by Michael McCloskey**
Newton's laws are well known, but tests show many people think moving objects act otherwise.

DEPARTMENTS

6 LETTERS
11 50 AND 100 YEARS AGO
12 THE AUTHORS
14 METAMAGICAL THEMAS
30 BOOKS
69A SCIENCE AND THE CITIZEN
132 THE AMATEUR SCIENTIST
138 BIBLIOGRAPHY

BOARD OF EDITORS Gerard Piel (Publisher), Dennis Flanagan (Editor), Brian P. Hayes (Associate Editor), Philip Morrison (Book Editor), John M. Benditt, Peter G. Brown, Michael Feirtag, Diana Lutz, Jonathan B. Piel, John Purcell, James T. Rogers, Armand Schwab, Jr., Joseph Wisnovsky

ART DEPARTMENT Samuel L. Howard (Art Director), Steven R. Black (Assistant Art Director), Ilil Arbel, Edward Bell

PRODUCTION DEPARTMENT Richard Sasso (Production Manager), Carol Hansen and Leo J. Petruzzi (Assistants to the Production Manager), Carol Eisler (Senior Production Associate), Karen O'Connor (Assistant Production Manager), Carol Albert, Lori Mogol, Martin O. K. Paul, Julio E. Xavier

COPY DEPARTMENT Sally Porter Jenks (Copy Chief), Debra Q. Bennett, Mary Knight, Dorothy R. Patterson

GENERAL MANAGER George S. Conn

ADVERTISING DIRECTOR C. John Kirby

CIRCULATION MANAGER William H. Yokel

SECRETARY Arlene Wright

SCIENTIFIC AMERICAN (ISSN 0036-8733), PUBLISHED MONTHLY BY SCIENTIFIC AMERICAN, INC., 415 MADISON AVENUE, NEW YORK, N.Y. 10017. COPYRIGHT © 1983 BY SCIENTIFIC AMERICAN, INC. ALL RIGHTS RESERVED. PRINTED IN THE U.S.A. NO PART OF THIS ISSUE MAY BE REPRODUCED BY ANY MECHANICAL, PHOTOGRAPHIC OR ELECTRONIC PROCESS, OR IN THE FORM OF A PHONOGRAPHIC RECORDING, NOR MAY IT BE STORED IN A RETRIEVAL SYSTEM, TRANSMITTED OR OTHERWISE COPIED FOR PUBLIC OR PRIVATE USE WITHOUT WRITTEN PERMISSION OF THE PUBLISHER. SECOND-CLASS POSTAGE PAID AT NEW YORK, N.Y., AND AT ADDITIONAL MAILING OFFICES. AUTHORIZED AS SECOND-CLASS MAIL BY THE POST OFFICE DEPARTMENT, OTTAWA, CANADA, AND FOR PAYMENT OF POSTAGE IN CASH. SUBSCRIPTION RATE: $21 PER YEAR IN THE U.S. AND ITS POSSESSIONS; $27 PER YEAR IN ALL OTHER COUNTRIES. POSTMASTER: SEND ADDRESS CHANGES TO SCIENTIFIC AMERICAN, 415 MADISON AVENUE, NEW YORK, N.Y. 10017.

SCIENTIFIC AMERICAN

CORRESPONDENCE

Offprints of more than 1,000 selected articles from earlier issues of this magazine, listed in an annual catalogue, are available at 75 cents each. Correspondence, orders and requests for the catalogue should be addressed to W. H. Freeman and Company, 4419 West 1980 South, Salt Lake City, UT 84121. Offprints adopted for classroom use may be ordered direct or through a college bookstore. Sets of 10 or more Offprints are collated by the publisher and are delivered as sets to bookstores.

Photocopying rights are hereby granted by Scientific American, Inc., to libraries and others registered with the Copyright Clearance Center (CCC) to photocopy articles in this issue of SCIENTIFIC AMERICAN for the flat fee of $2 per copy of each article or any part thereof. Such clearance does not extend to the photocopying of articles for promotion or other commercial purposes. Correspondence and payment should be addressed to Copyright Clearance Center, Inc., 21 Congress Street, Salem, MA 01970. Specify CCC Reference Number ISSN 0036–8733/83. $2.00 + 0.00.

Editorial correspondence should be addressed to The Editors, SCIENTIFIC AMERICAN, 415 Madison Avenue, New York, NY 10017. Manuscripts are submitted at the authors' risk and will not be returned unless they are accompanied by postage.

Advertising correspondence should be addressed to C. John Kirby, Advertising Director, SCIENTIFIC AMERICAN, 415 Madison Avenue, New York, NY 10017.

Subscription correspondence should be addressed to Subscription Manager, SCIENTIFIC AMERICAN, P.O. Box 5969, New York, NY 10017. The date of the last issue on your subscription is shown in the upper right-hand corner of each month's mailing label. For change of address notify us at least four weeks in advance. Please send your old address (if convenient, on a mailing label of a recent issue) as well as the new one.

Name

New Address

Street

City

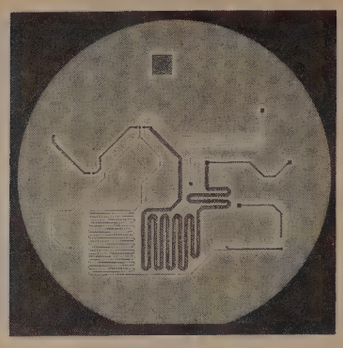
State and ZIP

Old Address

Street

City

State and ZIP



THE COVER

The painting on the cover illustrates a new technology based on "machining" silicon wafers like those used to make microelectronic chips (see "Silicon Micromechanical Devices," by James B. Angell, Stephen C. Terry and Phillip W. Barth, page 44). The cover painting shows a three-inch wafer that has been machined to make a gas chromatograph that will fit into the palm of the hand. The serpentine elements are gas channels formed by etching grooves into the silicon and bonding the wafer to a glass plate. Square holes etched completely through the wafer allow the gas to flow from system components etched into the front of the wafer to those mounted on the back. Underlying the double and triple sets of holes are valves; the valve seats are etched into the back of the wafer. The sample to be analyzed in the device is injected through a valve into the widest serpentine channel, where it is compressed by a tiny piston. The pressure in the channel is monitored by a silicon sensor mounted on the wafer. On command the pressurized sample is injected into both a lined separation column (such as a hollow fiber of fused silica) and an unlined reference column (the serpentine channel of intermediate diameter on the front of the wafer). Both the separation and the reference columns lead to a gas channel etched into the back of the wafer, over which is mounted a thermal-conductivity detector on a silicon chip. A portable gas analyzer consisting of five such chromatographs and a microcomputer is manufactured by Microsensor Technology, Inc.

THE ILLUSTRATIONS

Cover painting by George V. Kelvin

Page	Source
14–20	Edward Bell
38–39	Ilil Arbel
40–41	U.S. Army
42–43	Sovfoto
45	Stanford University Integrated Circuits Laboratory, courtesy of Phillip W. Barth
46–49	Gabor Kiss
50	Stanford University Integrated Circuits Laboratory, courtesy of Rosemarie Koch
51–52	Gabor Kiss
53–54	Stanford University Integrated Circuits Laboratory, courtesy of Phillip W. Barth (*top*); Gabor Kiss (*bottom*)
55	IBM San Jose Research Laboratory, courtesy of Kurt E. Petersen
56–68	Jerome Kuhl
79–81	Woods Hole Oceanographic Institution
82–92	Ian Worpole
95–96	Tom Prentiss
97	Tom Prentiss (*left*), Adrian R. Morrison (*right*)
98	Adrian R. Morrison
99–102	Tom Prentiss
105	Ryszard Grygiel
106	Ryszard Grygiel (*top*), Peter Bogucki (*bottom*)
107–112	Alan D. Iselin
114–116	Patricia J. Wynne
117	Kenneth B. Raper
118	Patricia J. Wynne (*top*), John Tyler Bonner (*bottom*)
119	Patricia J. Wynne
120	Kenneth B. Raper
123	Michael Goodman
124	Fowler Architectural Collection, Special Collections Division, Milton S. Eisenhower Library, Johns Hopkins University
125–126	Michael Goodman
127	Biblioteca Nazionale Centrale, Florence
128–130	Michael Goodman
133	Matthew G. Wheeler
134–136	Michael Goodman

Pan Am's Clipper Class Seat 18½"

TWA's Ambassador Class Seat 20⅞"

British Airways Super Club Seat 24"*

THE WORLD'S WIDEST AIRLINE SEAT.

Who else but British Airways.

Who else indeed! After all, no other airline on earth offers you a business class in a class with British Airways' Super Club.® The world's roomiest business class seat, always next to an aisle or a window, and service as only the British can provide it.

Should you choose to step up a bit, you'll find our Crown First Class™ waiting, with a sleeperseat of noble proportions. It seems a shame to sleep through our haute cuisine and fine china service, but it's nice to be able to.

Finally, there's the stratosphere of luxury, our Concorde. The only supersonic flights to Britain.

And, you'll be pleased to note that your flight miles between the US and London on British Airways will count as credit toward the A Advantage℠ travel award plan.

But, it's no more than you'd expect from the airline that flies more people to more countries than any other airline. Call your travel agent or British Airways.

British airways

*British Airways Super Club has a few seats 22" wide because of structural requirements, however, all Super Club Seats are wider than our competitors'.

SAVE AT ELEK-TEK

HEWLETT-PACKARD

It's Here

HP-75C
16K PORTABLE COMPUTER
$745.00 Mfr Sugg 995

The battery powered portable computer which includes a built-in HP-IL interface, card reader, and time and appointment modes. Typewriter keyboard.

HP-IL Products also available:

	Mfr Sugg	Elek-Tek
Digital Cassette Drive	450	$349
Thermal Printer/Plotter	450	349
Video Interface	225	175
Solution Books	35	29

CALCULATOR SAVINGS

	Mfr Sugg	Elek-Tek
HP-41C	195	$149
HP-41CV	275	209
Optical Wand	125	95
Card Reader	195	149
Printer	385	283
Quad R.A.M.(for HP41C)	75	60
Ext. Memory Module	75	60
Ext. Function Module	75	60
Time Module	75	60

HEWLETT-PACKARD
SLIMLINE Programmable
LCD CALCULATORS

HP-10C Scientific $54
HP-11C Scientific 70
HP-15C Scientific 92
HP-12C Financial 92
HP-16C Programmer 92

LIMITED TIME OFFER!
HP-11C, 12C, 15C
Solution Books 8.50
purchased with calculator.

Special! 50% off mfr. suggested retail price ($150.00 ea.)
HP-34C Scientific 75
HP-38C Financial 75

PERSONAL COMPUTING DEALER

CALL TOLL FREE 800-621-1269
EXCEPT Illinois, Alaska, Hawaii

Access. discounted too. Corporate Accts. Invited. Mastercard or Visa by mail or phone. Mail Cashier's Check, Money Ord. Pers. Check (2wks to clr.) Add $4.00 1st item. (AK, HI, P.R., Canada add $8.00 first item) $1.00 ea. add'l shpg. & handl. Shipments to IL address add 6% tax. Prices subj. to change. WRITE for free catalog. ALL ELEK-TEK MERCHANDISE IS BRAND NEW, FIRST QUALITY AND COMPLETE.

ELEK-TEK, inc.
6557 N. Lincoln Ave., Chicago IL 60645
(800) 621-1269 (312) 677-7660

LETTERS

Sirs:

I was very interested to read Douglas R. Hofstadter's article on nonsense ["Metamagical Themas," SCIENTIFIC AMERICAN, December, 1982], both as a professional mathematician and as a connoisseur of nonsense verse myself. I feel, however, that he omitted two further significant sources of nonsense.

One is the works of the English playwright Tom Stoppard. For example, in *Dogg's Hamlet—Cahoot's Macbeth* (in a context that I cannot fully describe here) Stoppard is rendering part of Act V, scene 8 of *Macbeth* into "Dogg" language thus:

"Rafters Birnam cakehops hobble Dunsinane, fry counterpane nit crossly window-framed, fancifully oblong! Sundry cobbles rattling up so chamberlain. Frantic, Macduff! Fry butter ban loss underlay—November glove!"

Secondly, I draw your attention to computer-generated nonsense. I enclose a sample of my own production. The sonnet "How Can the Purple Yeti Be So Red?" gained some fame in this country when it appeared in *The Times* as part of a discussion on whether computers could write verse. I think computers have a useful role to play here, both in randomly juxtaposing unlikely words and in generating new ones.

How can the purple yeti be so red,
Or chestnuts, like a widgeon,
  calmly groan?
No sheep is quite as crooked as a bed,
Though chickens ever try
  to hide a bone.
I grieve that greasy turnips
  slowly march:
Indeed, inflated is the icy pig:
For as the alligator strikes the larch,
So sighs the grazing goldfish for a wig.
Oh, has the pilchard argued with a top?
Say never that the parsnip is too weird!
I tell thee that a wolf-man will not hop
And no man ever praised the convex
  beard.
Effulgent is the day when bishops turn:
So let not then the doctor wake the urn!

Oh anaconda, tell me why the crane
Should be samoan when the toadstools
  scream:
A wailing hermit never maims a brain,
Although 'tis true that felons harm
  a bream.
My heart is spacious, likewise is it red,
When e'er I see the crazy carrots write;
I lost the briny princess—for a bed
Had madly spluttered as it chewed a
  light.
Alas! the days of midwife, elk and bat
Are gone, and now the hungry
  bailiffs blink;
Momentous was the crocus, now so fat
And ospreys cannot squash the
  smiling drink.
I shall no longer hide the ancient goose:
Life's not an ogre, but a gruesome
  moose!

J. R. PARTINGTON

Pembroke College
Cambridge, England

Sirs:

I read Jearl Walker's "Amateur Scientist" column in January with particular interest because I have a special interest in historical measurements of the universal constant of gravitation. There are two details in his résumé of Henry Cavendish's experiment that I should like to correct.

In the first place the torsion-balance experiment was planned by the Reverend John Michell, who unfortunately died before he could use the apparatus he had developed. It was this apparatus that Henry Cavendish took over, and although he reconstructed it, he kept close to Michell's plan and dimensions.

Less recherché, and more important, is the fact that Cavendish did not use an optical lever ("a light beam reflected from a small mirror") to measure the angle of rotation. Instead he observed vernier scales fitted to the ends of the torsion rod that moved against fixed ivory scales. These scales were illuminated by light focused on them from lamps fixed in wall apertures, and telescopes similarly fixed were used to view the scales.

G. S. LEADSTONE

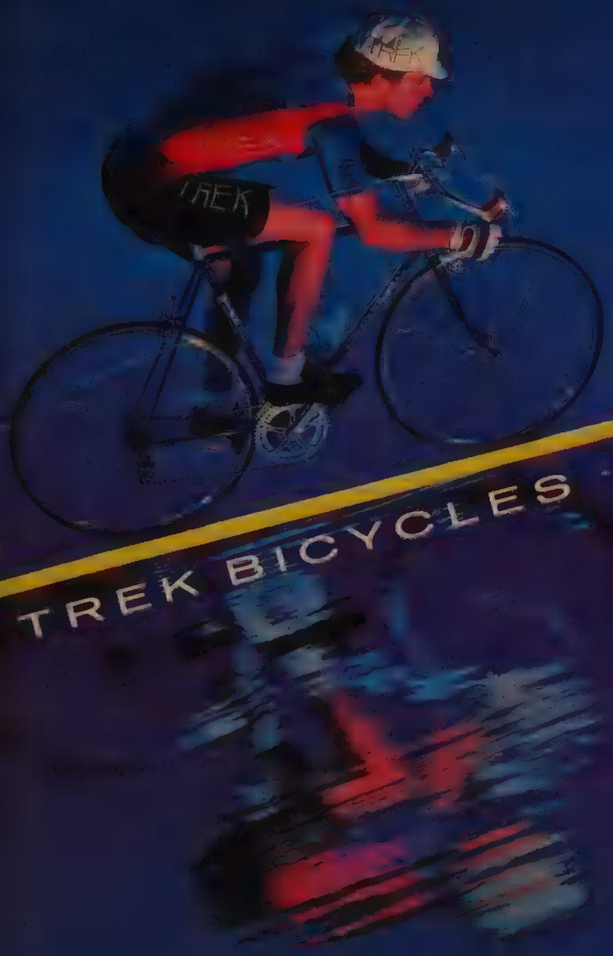
Atlantic College
Glamorgan, South Wales

Sirs:

"The Physics of Organ Pipes," by Neville H. Fletcher and Suszanne Thwaites [SCIENTIFIC AMERICAN, January], mentions the "tin-rich alloy" used for diapasons. The late organist and choirmaster Edward Whitney Flint of the Brooks School in North Andover, Mass., told the story that during the London blitz of World War II, the organ and the stained-glass windows of St. Paul's Cathedral were dismantled for safekeeping. After the war the organ was reassembled, except for a few missing pipes. One day the organist happened to be in the upper reaches of the cathedral when he came on a workman. On the scaffold beside him, being melted into solder for replacing the windows, was one of the missing pipes.

SAMUEL R. PHILLIPS

Portola Valley, Calif.



TREK BICYCLES · BRILLIANTLY RESPONSIVE.

The first criterion in defining bicycle performance is frame design. At Trek, we take bicycle frames seriously.

Designed by craftsmen who forged their reputations building custom framesets, every Trek bicycle frame is tailored to a specific riding need. This Function-Specific™ design redefines the relationship of rider to machine, providing the crisp, exhilarating experience of total response.

Trek frames are crafted of high strength-to-weight chrome or manganese molybdenum double-butted tubing.

Trek then matches components from around the world to complement the Function-Specific™ design of these outstanding frames.

The result—uncommon bicycles of exceptional strength and resiliency. A synthesis of meticulous design, esoteric materials and American craftsmanship.

Trek bicycles transcend the commonplace.

For the name of your nearest Trek dealer call Toll Free 800-558-TREK Ext. 400 (In Wisconsin call 414-478-2191)

TREK®

AMERICAN HANDBUILT BICYCLES

TREK BICYCLE CORPORATION
WATERLOO, WISCONSIN 53594

Authorized Trek Dealers

This is a partial listing of Trek dealers. If there is no dealer in your area, call 800-558-TREK. (In Wisconsin call 414-478-2195.)

Alabama
Auburn/Freewheeler 887-8117
Birmingham/Cahaba Cycles 969-0476

Arizona
Phoenix/Happiness House Schwinn Cyclery 995-8474
Phoenix/Landis Cyclery 264-5681
Tempe/Landis Cyclery 839-9383
Tucson/Bicycles West 887-7770

California
Anaheim/Anaheim Schwinn Cyclery 535-3231
Berkeley/Missing Link 843-7471
Bonita/Bonita Schwinn Cyclery 475-7205
Camarillo/Open Air Bicycles 484-4433
Claremont/Bud's Bike Shop 626-3285
Fullerton/Fullerton Bike Shop 879-8310
Goleta/Open Air Bicycles 967-9669
Huntington Beach/Two Wheel Transit Authority 536-6940
Isla Vista/Open Air Bicycles 968-5571
Laguna Beach/Laguna Beach Cyclery 494-1522
Los Altos/Bicycle Outfitter 948-8092
Los Angeles/I. Martin Imports Bicycles 653-6900
Northridge/Cycle World 349-6900
Palo Alto/Palo Alto Bikes 328-7411
Port Hueneme/Open Air Bicycles 985-5045
Redwood City/Chain Reaction Bicycles 366-7130
Riverside/Bicycle Bill's 689-5461
Sacramento/City Sport Works 332-6453
San Francisco/Presidio Bicycle Shop 751-3200
Santa Barbara/Open Air Bicycles 963-3717
Santa Rosa/The Cycle Path Ltd. 526-7121
Ventura/Open Air Bicycles 653-1100
Walnut Creek/Encina Bicycle Center 938-8943

Colorado
Aspen/Hub of Aspen 925-7970
Boulder/High Wheeler 442-5588
Colorado Springs/Criterium 475-0149
Denver/Denver Spoke 777-1720
Denver/Life Cycle 572-8405
Denver/Turin Bicycles, Ltd. 837-1857
Durango/Four Faces 247-0500
Estes Park/Cosmic Wheels Cyclery 586-2975
Fort Collins/The Bike Broker 484-2453

Connecticut
Bloomfield/Bloomfield Bike & Repair 242-9884
Old Greenwich/Buzz's Bicycle Shop 637-1665

Delaware
Newark/Two Wheeled Cycle 368-2685

Florida
Orlando/Orange Cycle Works 422-5552
South Miami/Bicycle Store of South Miami 665-0125

Georgia
Athens/Dixon's Bicycling Center, Inc. 549-2453
Atlanta/Vinings Schwinn 952-7731
Dunwoody/Dunwoody Schwinn 455-3171

Idaho
Boise/George's Lightweight Cycles 343-3782

Illinois
Arlington Heights/Bikes Plus 398-1650
Chicago/Village Cycle 751-2412
Evanston/Turin Bicycle Co-op 864-7660
Niles/Spoke 'N Pedal 692-4240
Oak Park/Oak Park Cyclery 386-0282

Indiana
Bloomington/Bicycle Garage 339-3457
Fort Wayne/Denny's Bike Shop 432-7636
West Lafayette/Hodson's Bay Company 743-3506

Iowa
Davenport/World of Bikes 323-8227
Des Moines/Bike World 255-7047
Iowa City/International World of Bikes 351-8337

Kansas
Manhattan/Pathfinder 539-5639
Overland Park/The Bike Rack 642-6115
Wichita/Bicycle Peddler 262-8029

Louisiana
New Orleans/Bikesmith, Inc. 891-2453
New Orleans/Gus Betat & Son 821-6386
Shreveport/Sunset Bicycle 635-5079

Maryland
Gaithersburg/Adam's Bicycles of Gaithersburg 840-0484

Massachusetts
Fall River/Paul's Schwinn 674-6343
Fitchburg/Gamache's Cyclery 343-3140
Worcester/O'Neil's Bicycle Shop 798-0084

Michigan
Ann Arbor/Great Lakes Cycling Center 663-1604
Kalamazoo/Alfred E. Bike 349-9423

Minnesota
Bloomington/Penn Cycle 888-1427
Hopkins/Now Sports 827-2838
Lake Elmo/North Country, Inc. 739-3500
Mankato/The Bike Shop 625-2453
Minneapolis/Freewheel Bike Co-op 339-2219
Richfield/Penn Cycle 866-7540
St. Paul/Now Sports 644-2354

Missouri
Bridgeton/Touring Cyclist 739-5183
Florissant/Touring Cyclist 921-1717
Maplewood/Maplewood Bicycle Sales & Service 781-9566
Richmond Heights/Touring Cyclist 781-7973

Montana
Billings/The Spoke Shop 656-8342

Nebraska
Lincoln/Cycle Tech, Inc. 466-2921
Omaha/Rock Brook Schwinn 391-3727

New Hampshire
Durham/Durham Bike 868-5634

New Jersey
Brielle/The Brielle Cyclery 528-9121
Montclair/Montclair Bikery 744-7252
Morristown/Marty's Reliable Cycle Center 538-7773
Ridgewood/Ridgewood Cycle Shop 444-2553

New Mexico
Albuquerque/Albuquerque Bike Co-op 265-5170
Los Alamos/Strings 'N Spokes 662-3456

New York
Brooklyn/Sheepshead Bicycle Shop 646-9430
Mt. Kisco/Bicycle World 666-4044
New York/Bicycle Habitat 691-2783
Rochester/Towner's 271-4453

North Carolina
Asheville/Liberty Corner Bicycles 253-7092

Ohio
Centerville/International Pro Bike Shop 433-6687
Columbus/Cross Country Cycle 294-4343
Findlay/The Bicycle Shop 423-2729
Yellow Springs/Yellow Springs Bike Shop 767-1956

Oklahoma
Oklahoma City/Pro Bike, Inc. 721-6707

Oregon
Eugene/Collins Cycle Shop 342-4878
Portland/The Bike Gallery 281-9800

Pennsylvania
Horsham/Pedal Pusher Ski & Sport 672-0202
State College/The Ski Station 237-2655

Texas
Austin/World Cycle 477-2142
Dallas/University Cycle 368-2035
Fort Worth/Bikes Etc. 927-2421
Lewisville/Lewisville Bicycle Center 221-2744

Utah
Salt Lake City/Joe Fisher's Bike Shop 466-3971
Salt Lake City/Sid's Ski & Cycle 485-3204

Virginia
Norfolk/Colley Avenue Bike Shop 622-0006

Washington
Seattle/Angle Lake Cyclery 878-7457
Seattle/Aurora Cycle 783-1000
Seattle/Northwest Cycle 329-2453
Spokane/Wheel Sport 326-3977

Wisconsin
Appleton/Janus 731-2253
Green Bay/Wheels & Boards 468-4745
Hales Corners/Wheel & Sprocket 425-7930
Madison/Yellow Jersey 257-4737
Milwaukee/Rainbow Jersey Bicycle Shop 961-1198
Monona/Village Pedaler 221-0311
Sun Prairie/Bicycle World 837-3440

Wyoming
Lander/Freewheel Sports 332-6616

Canada
Toronto/Bloor Cycle 536-9718

TREK®

COMPARED TO THE TOP THREE IMPORT SEDANS.

CHEVY CAVALIER PUTS YOU IN THE DRIVER'S SEAT. IN ROOM, POWER AND CHOICE.

Chevrolet Cavalier available after May 1, 1982.
Some Chevrolet cars are equipped with engines produced by other GM divisions, subsidiaries, or affiliated companies worldwide. See your dealer for details.

USA-1 IS TAKING CHARGE

CAVALIER · CELEBRITY · CITATION · CHEVETTE · CAMARO · MALIBU · MONTE CARLO · CAPRICE · CORVETTE

MORE POWER THAN THE TOP THREE IMPORTS.

Cavalier's new high-compression 2.0 Liter engine outpowers the three top-selling imports (based on SAE Net rated horsepower).

Coupled with Cavalier's front wheel drive and new-available 5-speed, it's performance you can feel.

Plus **Electronic Fuel Injection.** Chevy Cavalier has it. Accord doesn't. Corolla doesn't and Sentra doesn't.

MORE ROOM THAN THE TOP THREE IMPORT SEDANS.

The Cavalier Sedan shown below has 14 cu. ft. more total room than Toyota Corolla, 10 more than Nissan Sentra, 8 more than Honda Accord (based on EPA ratings for 1983 models).

Cavalier Coupe equals or beats Corolla, Sentra and Honda Prelude coupes in room. And Cavalier Wagon beats the top-selling import wagon by a comfortable margin—11 cu. ft. more room for passengers, 5 more for cargo.

NEW LOWER PRICES FOR 1983.

That's right. Based on a comparison of Manufacturer's Suggested Retail Prices for 1982 and 1983 models, Cavalier's prices are actually lower than last year. Level of equipment will vary.

AND YOU CAN CHOOSE FROM 5 MODELS -- INCLUDING A NEW CONVERTIBLE.*

Take your choice—there's room for five in the 4-Door Sedan, 2-Door Coupe, Hatchback Wagon and room for four in the exciting new Cavalier Convertible.

See your Chevy dealer now. Drive the proof that America's sales leader is taking charge.

GM

Chevrolet

SUN	MON	TUE	WED	THURS	FRI	SAT
1	2	3	4	5	6	7
8	9	10	11	12	13	14
15	16	17	18	19	20	21
22	23	24	25	26	27	28
29	30	31				

100% BLENDED SCOTCH WHISKIES, 86.8 PROOF.
IMPORTED BY SOMERSET IMPORTERS, LTD., N.Y.
© 1981

WOULDN'T IT BE WONDERFUL
IF EVERY BIRTHDAY
THIS MONTH
ENDED WITH RED.

GIVE JOHNNIE WALKER® RED
SEND IT ANYWHERE IN THE U.S.A.
CALL 1-800-528-6148

ITS DISTINCT TASTE MAKES IT THE WORLD'S BEST SELLING SCOTCH.

50 AND 100 YEARS AGO

SCIENTIFIC AMERICAN

APRIL, 1933: "The Italian scientist General G. A. Crocco suggests that for flight in the stratosphere at speeds above the speed of sound the internal-combustion engine must be replaced by a totally different type of prime mover, namely the jet-propulsion engine. He revives a suggestion of René Lorin, a Frenchman, who in 1913 proposed an engine that would take in the oncoming air, perhaps at a speed of 1,500 feet per second, through a front opening. The air would be highly compressed on entrance into the engine because of its great speed. Its temperature would rise, and at the same time it would mix with the fuel at 'auto-ignition' temperature; there would be combustion, a further increase in temperature and pressure, and a rush of air out at the rear of the engine at a speed greater than that of the aircraft as a whole. At high altitudes and at speeds of 1,000 miles per hour Lorin's mechanism can probably be made to give an efficiency just as high as that of the present-day combination of a propeller and an internal-combustion engine."

"The fundamental ideas of the behavioristic and Gestalt psychologies justify attempts to construct and develop machines of a new type—machines that can think. Such machines are entirely different from the integraphs, tide calculators and the like to which the term 'thinking machine' is sometimes applied. Unlike the latter they are not designed to perform with mathematical regularity but can 'learn' to vary their actions under certain conditions. Several mechanisms have recently been designed and built to illustrate the 'conditioned reflex,' regarded by behaviorists as the basic element of mental activity. These mechanisms exhibit in simplified form the type of activity studied by the Russian psychologist and physiologist Pavlov in his well-known experiments with dogs. Since it is possible to parallel the conditioned reflex as it is observed in the laboratory, the idea suggests itself that machines embodying this principle can be made to move about and to learn from stimuli encountered during their movement, just as animals do."

"The 'model' stars that astrophysicists use in their investigations must be imaginary—purely mathematical abstractions. We cannot get a sample of stellar material under stellar conditions. We might approximate the correct chemical composition, but pressures of billions of atmospheres and temperatures of billions of degrees are beyond our power to imitate. Each investigator has his own pet assumptions, and this is fortunate, for among the half dozen or so now active we may hope for a good reconnaissance of the possibilities. No one who is not eager for masses of computation should enter this field, for the calculations are laborious in the extreme. The stellar models that have so far been computed are curiously different. Some have central cores so dense that the ordinary gas laws fail and other equations must be used. Some are of enormous size; others are tiny—no larger than some of the planets. When the devoted labor of the computers has given us a large enough choice among them, we may hope to know more than we do now about the actual, as well as the possible, constitution of the stars."

APRIL, 1883: "By means of two dynamo-electric machines (Gramme system) of identically similar make M. Marcel Deprez has transmitted to Munich (over a distance of 57 kilometers) along an iron telegraph wire of 4.5 millimeters' diameter the power obtained at Miesbach from a steam engine. The receiving machine placed in the Crystal Palace at Munich supplied motive power for eight days to a centrifugal pump feeding a small waterfall about 2.5 meters in height. The success of the transmission of power from Miesbach to Munich is in every way an important event in the history of the technical applications of electricity."

"Dr. Roswell Park of Chicago describes the most recent applications of the electric light for surgical purposes. One form of these new instruments, called the gastroscope, is an instrument for the examination of the stomach. It consists of a bent tube with a window at one end, electric wires and tubes for the introduction of a water circulation to keep the tube cool while the electric light is burning. The tube is also equipped with reflector prisms and lenses for directing the light through it. A variety of other instruments have been made. For example, we have the laryngoscope, for examination of all parts of the throat; the oesophagoscope, for the gullet; the otoscope, for the ears; the urethroscope, for the bladder; the cystoscope, etc. They promise to be of utility and importance for the medical profession, for by their use many parts of the human system heretofore hidden from the eye may now be brilliantly lighted and examined and their condition in disease and health ascertained."

"Is there more than one force? We speak of the forces of nature and classify them as heat force, light force, electric force, etc. Should we not speak of the force of nature as exhibited in heat, in light, in electricity, etc.? You ask, if there is but one force, why does it not always manifest itself in the same way? Simply because the conditions are different at different times and in different places. The convertibility of these expressions of force, the one into the other, is perhaps the strongest proof of their identity of origin. It is impossible to study closely the properties of matter and the various phenomena exhibited in its relation to energy without coming to the conclusion that all forms of energy, however they may appear to our senses, are the offspring of one parent."

"Scientific prophets have foretold that a day will come when the 'residual products' resulting from distilling coal tar will be so valuable as to reduce the price of the gas obtained from the original coal to a mere nothing. A gentleman in England claims to know of an oven for making coke that enables those who use it to drive steam engines without any expense for fuel. Every ton of coal consumed in the oven yields coke worth 7s. and tar and ammonia worth 4s., in addition to 14,000 cubic feet of gas. If, therefore, the first two products are sold, the price—11s.—more than pays for the coal from which they were derived, as well as for labor, wear and tear, and interest on the capital sunk in plant. It is certainly a bold claim to put forward, but it is justified by the present prices of coke, tar and ammonia. If these ovens come into general use, the market value of such products will fall heavily in proportion to the immense enhancement of supply, and the prices fetched would not cover the cost of materials and labor."

"Running as an exercise can strengthen the limbs, develop the lungs, exercise the will and promote the circulation of the blood. The clothing should be light, the head bare and the neck uncovered. Care must be taken not to overdo. Running is well adapted to both young and middle-aged persons but not to those who are fat. Sedentary persons may find great benefit in it after the day's work is ended. If they live in cities, a quiet spot in the park may be selected and short trials adapted to the strength entered into. Girls can run as well as boys, and while they cannot go so fast, they can run much more gracefully. After puberty the change in the formation of the bones of the pelvis in girls renders running less easy and graceful. The modern style of dress for girls after puberty is also entirely unsuited to running."

THE AUTHORS

BARRY M. BLECHMAN and MARK R. MOORE ("A Nuclear-Weapon-Free Zone in Europe") are associated with the Roosevelt Center for American Policy Studies in Washington. Blechman is vice-president of the Roosevelt Center, responsible for its activities on foreign policy and defense policy. He received his undergraduate education at Queens College of the City University of New York. He then went on to obtain his master's degree in 1964 at New York University and his doctorate in international relations in 1971 from Georgetown University. From 1966 to 1971 he was on the staff of the Center for Naval Analyses, a research organization sponsored by the Federal Government. In 1971 he moved to the Brookings Institution, taking a leave in 1976 to become a member of President Carter's transition planning staff at the Office of Management and Budget. The following year he was appointed assistant director of the U.S. Arms Control and Disarmament Agency. He left the Government in 1979 to become a senior associate of the Carnegie Endowment for International Peace, moving from there to the Roosevelt Center. Blechman was a staff expert for the International Commission on Disarmament and Security Issues and helped to write its final report, in which the plan for a battlefield-nuclear-weapon-free zone in Europe was originally put forward. Moore is a graduate student at the Johns Hopkins School of Advanced International Studies and research assistant at the Roosevelt Center. He was graduated from Amherst College with a B.A. in 1981.

JAMES B. ANGELL, STEPHEN C. TERRY and PHILLIP W. BARTH ("Silicon Micromechanical Devices") are electrical engineers who did the work that forms the basis of their article when all three were at Stanford University. Angell is professor of electrical engineering and associate chairman of the department of electrical engineering at Stanford. He has three degrees from the Massachusetts Institute of Technology: an S.B. and an S.M. (both given in 1946) and an Sc.D. in electrical engineering (1952). From 1951 to 1960 he was a member of the research division of the Philco Corporation. At the end of that time he went to Stanford. An enthusiastic amateur musician, Angell is university carillonneur at Stanford, where he plays the 35-bell carillon in Hoover Tower. Terry is the founder of Microsensor Technology, Inc. His Ph.D. in electrical engineering (1975) is from Stanford, where he was senior research associate until he left in 1981 to found his company. Barth is senior research associate at the Stanford Integrated Circuits Laboratory. He was graduated from the University of Notre Dame with a B.S. in 1973 and went on to earn his M.S. (1974) and his Ph.D. in electrical engineering (1980) from Stanford. He is currently working with Angell on biomedical applications of miniature silicon sensors.

HAIM HARARI ("The Structure of Quarks and Leptons") is Annenberg Professor of High Energy Physics at the Weizmann Institute of Science. A fifth-generation Israeli, he received his master's degree and his doctorate in physics from the Hebrew University in Jerusalem. After serving for four years in the Israeli Army he joined the faculty of the Weizmann Institute in 1966. He is chairman of the planning and grants committee of the Council of Higher Education, the institution responsible for administering government financial support to the universities of Israel.

JOHN M. EDMOND and KAREN VON DAMM ("Hot Springs on the Ocean Floor") are marine geochemists. Edmond is professor of marine geochemistry at the Massachusetts Institute of Technology. He writes: "I am a native of Glasgow and got a degree in chemistry from the university there. Then I went to the Scripps Institution of Oceanography at the University of California at San Diego and did a Ph.D. on the carbon dioxide system in seawater. Following that I joined the faculty at M.I.T. The research efforts of my group there are divided about equally among three areas: the trace-metal geochemistry of seawater and general marine geochemistry; the chemistry of ridge-crest hot springs, and chemical processes in large lakes and big rivers and their estuaries. My outside interests include the politics, history and cultures of the overseas countries in which I have the opportunity to do field work. My current interest is in the linguistic development of my two-year-old Scots-Iranian son." Von Damm is a graduate student in the Joint Program in Oceanography conducted by M.I.T. and the Woods Hole Oceanographic Institution. She is finishing her doctoral thesis on the detailed chemical analysis of the waters vented by hot springs on the ocean floor.

ADRIAN R. MORRISON ("A Window on the Sleeping Brain") is professor of anatomy at the School of Veterinary Medicine at the University of Pennsylvania. He has two degrees from Cornell University: a D.V.M. (1960) and an M.S. (1962). His Ph.D. in anatomy (1964) is from the University of Pennsylvania. After a year of postdoctoral work at the University of Pisa he returned to Pennsylvania, where he has remained.

PETER BOGUCKI and RYSZARD GRYGIEL ("Early Farmers of the North European Plain") are archaeologists who have collaborated in work on the earliest farming cultures of north-central Poland since 1976. Bogucki is lecturer in anthropology at the University of Massachusetts at Boston. In the work with Grygiel his particular interest is the study of animal bones at the sites of the farming communities. He obtained his B.A. at the University of Pennsylvania in 1974 and went on to earn his Ph.D. in anthropology from Harvard University in 1981. In addition to his interest in prehistoric Europe he has been studying patterns of animal utilization by the colonial inhabitants of Portsmouth, N.H., and other New England towns. Grygiel is head of the Neolithic department at the Museum of Archaeology and Ethnography in Łódź. He earned his master's degree at Adam Mickiewicz University in Poznań in 1974 and his doctorate in archaeology from the University of Warsaw in 1980. His major scientific interest is the chronology of the early Neolithic communities in north-central Poland, and he has excavated many sites from the early Neolithic period.

JOHN TYLER BONNER ("Chemical Signals of Social Amoebae") is George M. Moffett Professor of Biology and chairman of the department of biology at Princeton University. He has three degrees from Harvard: a B.A. from Harvard College (1941) and an M.A. (1942) and a Ph.D. in biology (1947) from Harvard University. He writes: "In 1939 while an undergraduate I became interested in cellular slime molds (or social amoebae) as organisms ideally suited to experimental studies in developmental biology. It was a case of love at first sight and we have been going steady ever since. After graduate work and a stint in the Air Corps I came to Princeton, where I remain. Besides continuing my experimental work I have written a number of books, which mostly revolve around the general theme of evolution and development from cells to societies." Among Bonner's books are *Cells and Societies* (1955) and *On Development: The Biology of Form* (1977).

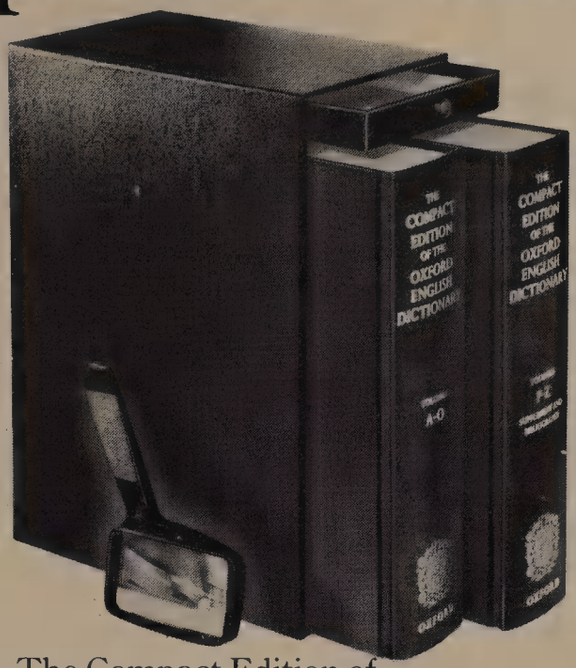
MICHAEL McCLOSKEY ("Intuitive Physics") is assistant professor of psychology at Johns Hopkins University. He was graduated from Emory University with a B.A. in 1975 and went on to get his Ph.D. in cognitive psychology from Princeton University in 1978 before moving to Johns Hopkins.

Important choices. Impressive savings.

The Encyclopedia of Philosophy

for **$24.95** (Pub price $175)

The most comprehensive encyclopedia of philosophy ever published, this set traces ancient, medieval, modern, Eastern and Western thought. An essential and rewarding reference source for home libraries.



The Compact Edition of The Oxford English Dictionary

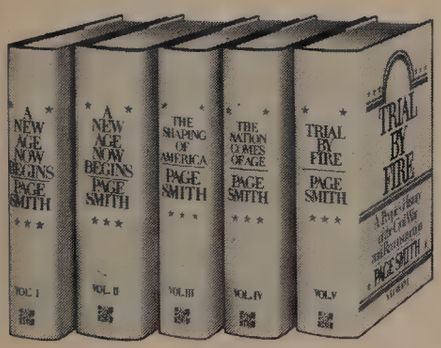
for **$24.95** (Pub price $150)

The "most complete, most scholarly dictionary of the English language"—*The Christian Science Monitor.* Through photo-reduction, the original 13-volume set has been reproduced in this two-volume *Compact Edition.* Magnifying glass included.

The Nine Beethoven Symphonies & The Five Piano Concertos/Ashkenazy

for **$19.95** (List prices total $115.74)

Two of the most acclaimed recordings of Beethoven's works. Sir Georg Solti conducts the Chicago Symphony in the nine-record set of Beethoven's complete symphonies. "A glorious musical combustion"—*Time.* In collaboration with soloist Vladimir Ashkenazy, Solti and the Chicago Symphony perform Beethoven's five concertos. "A major addition to the Beethoven discography"—*High Fidelity.* Available in 13 records or 8 cassettes.



Page Smith's People's History of the United States

for **$15.95** (Pub prices total $99.85)

A lively chronicle of America's people from Plymouth through the Civil War. A five-volume set written by "a historian whose scholarship flows from a storyteller's fingertips"—*Publishers Weekly.*

The Story of Civilization by Will and Ariel Durant

for **$29.95** (Pub prices total $335.45)

For almost half a century Will and Ariel Durant traced the continuity of world history—the religions and philosophies, the political and economic tides, the arts and sciences, the customs and conquests—to show the foundations of society today. A Book-of-the-Month Club exclusive for almost 50 years, the Durants' illustrated masterwork is history come alive.

Choose one of these five sets and save up to $305.50

You simply agree to buy 4 books within the next two years.

Facts About Membership. As a member you will receive the *Book-of-the-Month Club News*® 15 times a year (about every 3½ weeks). Every issue reviews a Selection and 150 other books that we call Alternates, which are carefully chosen by our editors. If you want the Selection, do nothing. It will be shipped to you automatically. If you want one or more Alternates—or no book at all—indicate your decision on the Reply Form and return it by the specified date. *Return Privilege*: If the *News* is delayed and you receive the Selection without having had 10 days to notify us, you may return it for credit at our expense. *Cancellations*: Membership may be discontinued, either by you or by the Club, at any time after you have bought four additional books. Join today. With savings and choices like these, no wonder Book-of-the-Month Club is America's Bookstore.

BOOK-OF-THE-MONTH CLUB

America's Bookstore® since 1926.

Book-of-the-Month Club, Inc., Camp Hill, Pennsylvania 17012 **A170-4**

Check one box only.

- ☐ 912. Compact OED **$24.95**
- ☐ 913. The Story of Civ. **$29.95**
- ☐ 917. Ency. of Philosophy **$24.95**
- ☐ 932. Beethoven (13 Records) **$19.95**
- ☐ 933. Beethoven (8 Cassettes) **$19.95**
- ☐ 954. Page Smith's History of U.S. **$15.95**

Please enroll me as a member of Book-of-the-Month Club and send me the set I have checked at left, billing me for the appropriate amount, plus shipping and handling charges. I agree to buy four books during the next two years. A shipping and handling charge is added to each shipment.

Mr.
Mrs. 3-64
Miss (Please print plainly)

Address Apt.

City

State Zip

Prices shown are U.S. prices. Outside the U.S., prices are generally higher.

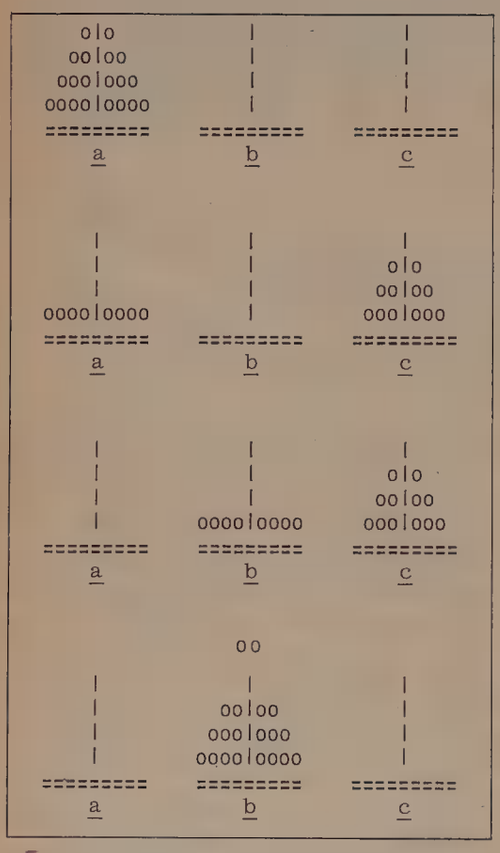
METAMAGICAL THEMAS

In which a discourse on the language Lisp concludes with a gargantuan Italian feast

by Douglas R. Hofstadter

Last month I described Édouard Lucas's Tower of Brahma puzzle, where the object is to transfer a tower of 64 gold disks from one diamond needle to another, making use of a third needle on which disks can be placed temporarily. The disks must be picked up and moved one at a time, the only other constraint being that no disk may ever be placed on a smaller one. The problem I posed for readers was to come up with a recursive description, expressed as a function in the computer language Lisp, of how to accomplish this goal (and thereby, according to the priests of Brahma, end the world).

I pointed out that the recursion is evident enough: in order to transfer 64 disks from one needle to another, using a third, it suffices to know how to transfer 63 disks from one needle to another, using a third. To recapitulate the idea, it is this. Suppose the 64-disk tower of Brahma starts out on needle *a*. At the top in the illustration on this page is a schematic picture representing all 64 disks by a mere 4. First of all, relying on your presumed 63-disk-moving ability, transfer 63 disks from needle *a* to needle *c*, using needle *b* as your "helping needle." How the setup looks is shown second from the top in the illustration. (Note that 4 plays the role of 64 in my original picture, so that 3 plays the role of 63, but for some reason 1 does not play the role of 61. Isn't that peculiar?) Now, simply pick up the one remaining *a* disk—the biggest disk of all—and plunk it down on needle *b*, as is shown third from the top.



The Tower of Brahma puzzle with four disks

Now you can see how easy it will be to finish up: simply reexploit your 63-disk ability to transfer that pile on *c* back to *b*, this time using *a* as the "helping needle." Notice how in this maneuver needle *a* plays the helping role needle *c* played in the preceding 63-disk maneuver. A split second before the last disk is put in place the situation is as it is shown at the bottom in the illustration. Why before it is in place? Why not after? Because the entire world then turns to dust, and it is too hard to draw dust.

Now, someone might complain that I left out all the hard parts: "You just magically assumed an ability to move 63 disks." So it may seem, but there is nothing magical about such an assumption. After all, in order to move 63 you merely need to know how to move 62. And to move 62, you merely need to know how to move 61. On it goes down the line, until you bottom out at the "embryonic case" of the Tower of Brahma puzzle, the 1-disk puzzle. Now, I admit that you have to keep track of where you are in the process, which may be a bit tedious, but it is merely bookkeeping. In principle you now could actually carry out the entire process—if you were bent on seeing the world end!

As our first approximation of a Lisp function let us write an English description of the method. Let us call the three needles "sn", "dn" and "hn", standing for source needle, destination needle and helping needle. Here goes:

```
To move a tower of height n
  from sn to dn, making use of hn:
    if n = 1,
      then just carry that one disk
      directly from sn to dn;
    otherwise,
      do the following three steps:
        (1) move a tower of height
        n − 1 from sn to hn,
          making use of dn;
        (2) carry 1 disk from sn to dn;
        (3) move a tower of height
        n − 1 from hn to dn,
          making use of sn.
```

Here the steps labeled 1 and 3 are the two recursive calls; skirting paradox, they seem to call on the very ability they are helping to define. The saving feature is that they involve n − 1 disks instead of n. Note that in step 1 hn plays the "destination" role while dn plays the "helper" role. And in step 3 hn plays the "source" role while sn plays the "helper" role. Since the whole thing is recursive, every needle will be switching hats many times over in the course of the transfers. That is the beauty of the puzzle and in a way it is also the beauty of recursion.

How do we make the transition from English to Lisp? It is quite simple:

```
(def move-tower
 (lambda (n sn dn hn)
 (cond
  ((eq n 1) (carry-one-disk sn dn))
  (t (move-tower (sub1 n) sn hn dn)
    (carry-one-disk sn dn)
    (move-tower (sub1 n) hn dn sn)))))
```

Where are the Lisp equivalents of the English words "from," "to," and "making use of"? They seem to have disappeared. Then how can the Lisp genie know which needle is to play which role at each stage? The answer is that the information is conveyed *positionally*. There are, in this function definition, four parameters: one integer and three "dummy needles." The first of the three needles is the source, the second the destination and the third the helper. Thus in the initial list of parameters (following the "lambda") they are in the order "sn dn hn". In the first recursive call, the Lisp translation of step 1, they are in the order "sn hn dn", showing how hn and dn have switched hats. In the second recursive call, the Lisp translation of step 3, you can see that hn and sn have switched hats.

The point is that the atom names "sn", "dn" and "hn" carry no intrinsic meaning to the genie. They could as well have been "apple", "banana" and "cherry". Their meanings are defined operationally, by the places where they appear in the various parts of the function defini-

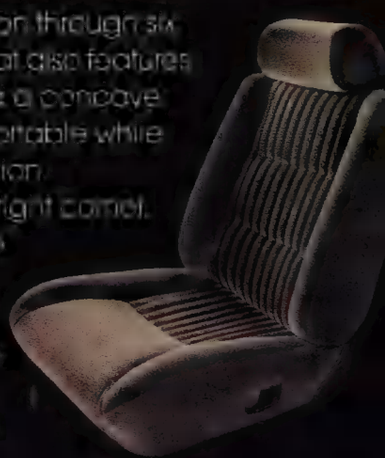
ONLY PONTIAC HAS IT. OR ANYTHING LIKE IT.

ONTIAC 6000 STE

s is the Pontiac with more technologically advanced atures than any other. And the future of American luxury formance sedans. As the 3.1 liter high output V-6 sparks to and the graphic instrumentation lights up, you gain a cise driver's sense of control over your environment. ou are truly in touch with your car. And the STE's Driver ormation Center is helping to put there. Important engine functions, p operation, door, hood and trunk curity – even service reminders – can be monitored from the driver's seat with a simple touch of your finger. You can adjust the climate of your driver environment. It's easy with the STE's new climate control center which uses advanced electroluminescent and light-emitting diode readout.

Now notice the seat you're in. It's biomechanically engineered for serious drivers, and can be fine-tuned to your individual driving position through six areas of adjustment. The STE seat also features thigh and lumbar support and a concave shape to help keep you comfortable while you take the curves with precision.

When you cut a particularly tight corner, you won't be thinking about the self-leveling Electronic Ride Control. The exclusive "pore free" cast aluminum wheels. Or the Goodyear Eagle GT high performance radials with analytically tuned shock absorbers, springs, bushings, front and rear stabilizer bars, power rack and pinion steering. Oh how these components perform harmoniously as one system. You'll just be amazed at how the responsive Pontiac 6000 STE suspension works as an extension of yourself.



If you're used to traditional luxury sedans, we think you'll find this experience refreshing. And very exciting.

SPECIAL EDITION STE SPECIAL TOURING EDITION

are

for 1984.

GM

PONTIAC WE BUILD EXCITEMENT

How to turn data you can see.

There are more people in more places doing more things with Apples than with any other personal computer in the world.

Which is saying a lot.

But we'd like to take the time to explain just one of the things that can make an Apple® Personal Computer more meaningful to you, personally.

It's called business graphics.

An important business tool that demonstrates, quite graphically, how one picture is worth a few thousand numbers.

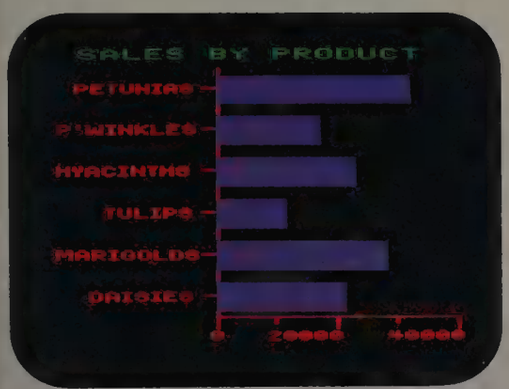
More ways to see what you're doing.

That's the beauty of Apple Business Graphics. The software package that gives your Apple II or Apple III personal computer the power to convert rows and rows of numbers—

a sea of data into

or pounds of printout—into one simple, colorful, comprehensible illustration.

And that gives you the power, for example, to thoroughly understand relationships and comparisons in a complex market analysis. Without having to spend hours sifting through



Horizontal bar graph (a unique feature) makes it easy to compare the profitability of six different products—from periwinkles to petunias.

dull data—armed with little more than a ruler, a compass and a few not-so-magic markers.

With Apple Business Graphics, all you do is enter the data or transfer it directly from your VisiCalc® model or a number of other computer files. Your Apple will take it from there.

Helping you to get the big picture in a variety of ways. Line or area graphs, bar graphs, pie charts or scattergrams.

Whichever configuration best suits your application.

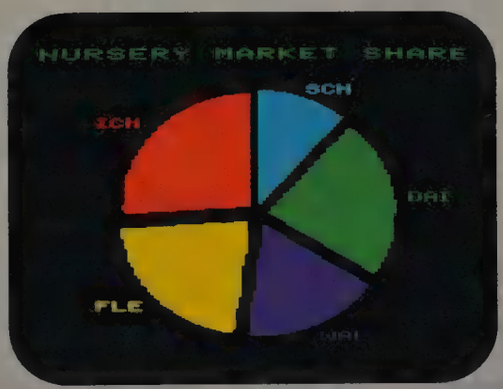
When you're through charting market shares you can do the same with forecasts, budgets, stock trends, business plans or customer demographics.

Show your stuff.

In addition to helping you better visualize results, Apple Business Graphics can make presentations more presentable.

Because whatever you've produced can be rapidly, easily and colorfully reproduced, on virtually any printer or plotter on the market.

Moreover, sophisticated Apple features like exploded views,



To get the overall market share picture, (and any other percentage related data) Ichiban management consults this easy-to-digest pie chart.

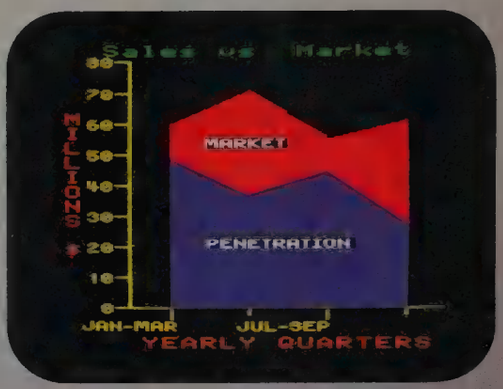
unlimited overlays and floating titles allow your displays to be remarkably flexible.

By combining Apple Business Graphics with various slide-show packages, charts and graphs on your monitor can easily be converted into impressive slides and transparencies.

Getting started in pictures.

Start by seeing any of the over 1300 authorized Apple dealers.

They'll show you Apple Business Graphics software can generate more types of pictures, in more colors, using more data



With up to six colors per graph, employees can easily see just how their annual sales blossomed relative to the rest of the market.

than any other graphics package.

More important, they'll show you what all that can mean to you, your accounting firm, your import business, or your chain of quiche-on-a-bun family restaurants.

They'll even help you produce your first picture.

But business graphics is just one of the things an Apple can help you do. One of thousands. Which is not terribly surprising, when you consider there's more software available for Apples than for any other personal computer.

So consider a trip to your Apple dealer. And draw your own conclusions.

apple®

The most personal computer.

Call (800) 538-9696 for the location of the authorized Apple dealer nearest you, or for information regarding corporate purchases through our National Account Program. In Calif. (800) 662-9238. Or write Apple Computer Inc., Advertising and Promotion Dept., 20525 Mariani Ave., Cupertino, CA 95014. VisiCalc is a registered trademark of VisiCorp.

© 1983 Apple Computer Inc.

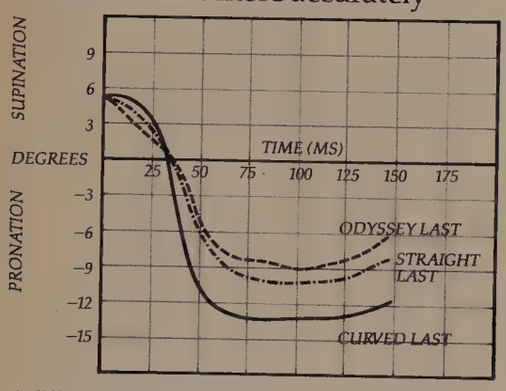
A NEW SHOE WILL NEVER FIT BETTER THAN THE LAST.

Let's get our terms straight.

From now on, when we use the word "last," we're talking about what comes first.

We're talking about that rather strange looking device over which the shoe is initially formed. Now this last can be made of wood, metal or plastic. But what's truly critical, what keeps us up nights, is its shape.

In fact, at Nike, we think the shape of the last is so important that we've launched one of the most comprehensive foot morphology studies since the 1940's. Because the only way to give you a better fitting shoe is to come up with a last that more accurately



A different last means different degrees of rearfoot control. Data obtained from high speed films of 10 subjects running at 7:00 mile pace in shoes based on three different lasts.

reflects both the structure and dynamics of your individual foot.

It's a tough job. And we're far from done. But with the data we've already collected, combined with extensive fit trials and wear tests, we've been able to make some radical changes. Changes that not only affect the comfort of Nike shoes, but more surprisingly, also improve their inherent stability.

Curved Last *Straight Last* *Odyssey Last*

The normal foot, as seen in each of three new Nike lasts. The Odyssey provides the most support under the medial arch.

Today, there are three new Nike lasts. With each, we widened the heel base, trimmed the Achilles area and put what amounts to a slight curve in the heel area. As a result, the foot is now centered more directly over the middle of the shoe.

Which, as tests prove, effectively reduces rearfoot motion.

Up front, we gave the big toe additional room. Both overhead and on the medial side. This too promotes stability and may even have a beneficial influence on the foot's ability to create its own cushion.

Our recommendations:

If you're a fast runner or have high-arched rigid feet, try a model from our new curved last. Namely, the Columbia, Aurora, Valkyrie, Internationalist, Pegasus, or the Elite Classic.

On the other hand, if you run a bit slower, or have low-arched flexible feet, or pronate to excess (that is, roll to the inside of your shoes), slip into the new Equator made on our new straight last. It does an even better job of controlling rearfoot motion because of the increased support under the medial arch.

Our most revolutionary last, however, we dubbed the Odyssey. Then we built a shoe of the same name to go with it. As a last, it's the straightest of the bunch. And also the most versatile. Runners who pronate a lot, receive a lot of correction. Those who pronate less, get less. And just about everyone finds it exceedingly comfortable.

The Odyssey

The Equator

Unfortunately, as important as a last may be, you'll probably never see one. They simply never leave our factory.

It's frustrating, we know. But then again, sometimes what helps you the most is the thing that gives you fits.

NIKE

Beaverton, Oregon

n. Hence it would have been a gross under to have written, for instance, move-tower (sub1 n) sn dn hn)" as sp for step 1, because that contains no dication that hn and dn must switch les in that step.

An important question remains. What ppens when the friendly Lisp genie mes to a step that says "carry-one-sk"? Does it suddenly zoom off to the mple at Benares and literally heft a lid gold disk? Or more prosaically, es it pick up a plastic disk on a table d transfer it from one plastic needle another? In other words, does some ysical action, rather than a mere com-itation, take place?

Well, in theory it is quite possible. In ct, even in practice the execution of a isp command could actually cause a echanical arm to move to a specific cation, to pick up whatever its me-anical hand grasps, to carry that ob-ct to another specific location and then release it. In these days of industrial bots there is nothing science-fictional bout it. In the absence of a mechanical rm and hand to move physical disks, owever, what could it mean?

One obvious and closely related possi-ility is for there to be a display of the uzzle on a screen, and for "carry-one-isk" to cause a *picture* of a hand to ove, to *appear* to grasp a disk, pick it p and replace it somewhere else. That ould amount to simulating the puzzle raphically, which can be done with arying degrees of realism.

Suppose, however, we do not have ancy graphics hardware or software at ur disposition. Suppose all we want is o create a printed recipe telling us how o move our soft, organic, human hands n order to solve the puzzle. Thus in the ase of a three-disk puzzle the desired ecipe might read "ab ac bc ab ca cb b", or equally well "12 13 23 12 31 2 12". What could help us to reach his humbler goal?

If we had a program that moved an rm, we would be concerned not with he value it returned but with the pat-erned sequence of "side effects" it car-ied out. Here, in contrast, we are con-erned only with the value our program s going to return: a patterned list of at-oms. The list for n = 3 has got to be built up from two lists for n = 2. This idea was shown at the end of last month's column:

ab ac bc ab ca cb ab

In order to set groups apart in Lisp, rather than using wider spaces we use parentheses. Hence our goal for n = 3 might be to produce the sandwichlike list "((ab ac bc) ab (ca cb ab))". One way to produce a list out of its compo-nents is to use "cons" repeatedly. There-fore if the values of the atoms "apple", "banana" and "cherry" are 1, 2 and 3,

```
(ab ac bc ab ca cb ab ac bc ba ca bc ab ac bc ab ca cb ab ca bc ba ca cb
 ab ac bc ab ca cb ab ac bc ba ca bc ab ac bc ba ca cb ab ca bc ba ca bc
 ab ac bc ab ca cb ab ac bc ba ca bc ab ac bc ab ca cb ab ca bc ba ca cb
 ab ac bc ab ca cb ab ca bc ba ca bc ab ac bc ba ca cb ab ca bc ba ca cb
 ab ac bc ab ca cb ab ac bc ba ca bc ab ac bc ab ca cb ab ca bc ba ca cb
 ab ac bc ab ca cb ab ac bc ba ca bc ab ac bc ba ca cb ab ca bc ba ca bc
 ab ac bc ab ca cb ab ac bc ba ca bc ab ac bc ba ca cb ab ca bc ba ca cb
 ab ac bc ab ca cb ab ca bc ba ca bc ab ac bc ba ca cb ab ca bc ba ca bc
 ab ac bc ab ca cb ab ac bc ba ca bc ab ac bc ab ca cb ab ca bc ba ca cb
 ab ac bc ab ca cb ab ac bc ba ca bc ab ac bc ba ca cb ab ca bc ba ca bc
 ab ac bc ab ca cb ab ca bc ba ca bc ab ac bc ba ca cb ab ca bc ba ca cb
 ab ac bc ab ca cb ab ac bc ba ca bc ab ac bc ab ca cb ab ca bc ba ca cb
 ab ac bc ab ca cb ab ca bc ba ca bc ab ac bc ba ca cb ab ca bc ba ca bc
 ab ac bc ab ca cb ab ac bc ba ca bc ab ac bc ba ca cb ab ca bc ba ca cb
 ab ac bc ab ca cb ab ca bc ba ca bc ab ac bc ba ca cb ab ca bc ba ca cb
 ab ac bc ab ca cb ab ac bc ba ca bc ab ac bc ab ca cb ab ca bc ba ca cb
 ab ac bc ab ca cb ab ac bc ba ca bc ab ac bc ba ca cb ab ca bc ba ca bc
 ab ac bc ab ca cb ab ac bc ba ca bc ab ac bc ab ca cb ab ca bc ba ca cb
 ab ac bc ab ca cb ab ca bc ba ca bc ab ac bc ba ca cb ab ca bc ba ca cb
 ab ac bc ab ca cb ab ac bc ba ca bc ab ac bc ba ca cb ab ca bc ba ca cb
 ab ac bc ab ca cb ab)
```

The Lisp genie fires back the moves to solve a 9-high Tower of Brahma puzzle

then the value returned by "(cons apple (cons banana (cons cherry nil)))" will be the list "(1 2 3)". There is, however, a shorter way to get the same result, namely to write "(list apple banana cherry)". It returns the same value. Similarly, if the atoms "sn" and "dn" are bound respectively to "a" and "b", then execution of the command "(list sn dn)" will return "(a b)". The function "list" is an unusual function in that it takes any number of arguments at all—even none, so that "(list)" returns the value of nil!

Now let us tackle the problem of what value we want the function called "carry-one-disk" to return. It has two parameters that represent needles, and ideally we should like it to return a single atom made out of those needles' names, such as "ab" or "12". For the moment it will be easier if we assume that the needle names are the numbers 1, 2 and 3. In this case to make the number 12 out of 1 and 2 it suffices to do a bit of arithmetic: multiply the first number by 10 and add on the second. Here is the Lisp for that:

```
(def carry-one-disk
 (lambda (sn dn)
  (plus (times 10 sn) dn)))
```

On the other hand, if the needle names are non-numeric atoms, we can use a standard Lisp function called "concat", which takes the values of its arguments (any number, as with "list") and concatenates them all to make one big atom. Thus "(concat 'con 'cat 'e 'nate)" returns the atom "concatenate". In such a case we could write:

```
(def carry-one-disk
 (lambda (sn dn) (concat sn dn)))
```

Either way we have solved the bottom-line half of the move-tower problem.

The other half of the problem is what the recursive part of move-tower will return. Well, that is pretty easy. We simply want it to return a sandwichlike list in which the values of the two recursive calls flank the value of the single call to carry-one-disk. And so we can modify our previous recursive definition very slightly, by adding the word "list":

```
(def move-tower
 (lambda (n sn dn hn)
  (cond
   ((eq n 1) (carry-one-disk sn dn))
   (t
    (list
     (move-tower (sub1 n) sn hn dn)
     (carry-one-disk sn dn)
     (move-tower (sub1 n) hn dn sn))))))
```

Now let us conduct a little exchange with the Lisp genie:

```
—> (move-tower 4 'a 'b 'c)
(((ac ab cb) ac (ba bc ac))
              ab
 ((cb ca ba) cb (ac ab cb)))
—>
```

Smashing! It actually works! Isn't that pretty? In last month's column this list was called "brahma".

Suppose we wanted to suppress all the inner parentheses, so that just a long, uninterrupted sequence of atoms would be printed out. For instance, we would get "(ac ab cb ac ba bc ac ab cb ca ba cb ac ab cb)" instead of the intricacy of brahma. It would be slightly less informative but more impressive in its opaqueness. In this case we would not want to use the function "list" to make our sandwich of three values but would have to use some other function that would remove the parentheses from the two flanking recursive values.

This is a case where the standard Lisp function "append" comes in handy. It splices any number of lists together, dropping their outermost parentheses

as it does so. Thus "(append '(a (b)) '(c) nil '(d e))" yields the five-element list "(a (b) c d e)" rather than the four-element list "((a (b)) (c) nil (d e))", which would be yielded if "list" rather than "append" had appeared in the function position. Using append and a slightly modified version of carry-one-disk to work with it, we can formulate a final definition of move-tower that does what we want:

```
(def move-tower
 (lambda (n sn dn hn)
  (cond
   ((eq n 1) (carry-one-disk sn dn))
   (t
    (append
     (move-tower (sub1 n) sn hn dn)
     (carry-one-disk sn dn)
     (move-tower (sub1 n) hn dn sn))))))

(def carry-one-disk
 (lambda (sn dn) (list (concat sn dn))))
```

To test this out I asked the Lisp genie to solve a 9-high Tower of Brahma puzzle. What it shot back at me, almost instantaneously, is at the top of the preceding page.

We have just been through a rather sophisticated and brain-taxing example of recursion. Now let us take a look at a recursion that offers us a different kind of challenge. This recursion comes from an offhand remark I made in the last column. I used the odd variable name "tato", mentioning that it is a recursive acronym standing for "tato (and tato only)". Using this fact you can expand "tato" any number of times. The sole principle is that each occurrence of "tato" on a given level is replaced by the two-part phrase "tato (and tato only)" to make the next level. How it goes is shown in the illustration below. For us the challenge is to write a Lisp function that returns the value of "tato" after n recursive expansions, for any n. It is irrelevant that for n much bigger than 3 the whole thing gets ridiculously large. We are theoreticians!

There is only one problem: any Lisp function must return a single Lisp structure (an atom or a list) as its value, and the entries in our table do not satisfy this criterion. For instance, the entry for n = 2 consists of one atom followed by two lists. To fix this we can turn each entry in the table into a list by enclosing it in one outermost pair of parentheses. Now our goal is consistent with Lisp. How do we attain it?

Recursive thinking tells us that the bottom line, or embryonic case, occurs when n equals 0, and that otherwise the nth line is made from the line before it by replacing the atom "tato", wherever it occurs, with the list "(tato (and tato only))", but without its outermost parentheses. We can write this up:

```
(def tato-expansion
 (lambda (n)
  (cond
   ((eq n 0) '(tato))
   (t
    (replace
       'tato
       '(tato (and tato only))
       (tato-expansion (sub1 n)))))))
```

The only thing is that we have not specified what we mean by "replace". We must be very careful in defining how to carry out our "replace" operation. Look at any line of the "tato" table and you will see that it contains one more element than the preceding line. Why is this? Because the atom "tato" gets replaced each time by a two-element list whose parentheses, as I pointed out above, are dropped in the act of replacement. It is this parenthesis dropping that is the sticky point. A less tricky example of such parenthesis-dropping replacement than the recursive one involving "tato" would be "(replace 'a '(1 2 3) '(a b a))", whose value should be "(1 2 3 b 1 2 3)". Rather than exact substitution of a list for an atom, this kind of replacement involves splicing or appending inside a long list.

Let us try to specify in Lisp—using recursion as usual—just what we mean by "replacing" all occurrences of the atom "atm" by a list called "lyst", inside a long list called "longlist". This is a good puzzle for you to try. Let me give you a hint: See how the answer for argument "(a b a)" is built out of the answer for argument "(b a)". Also look at other simple cases like that, moving back down toward the embryonic case.

```
n=0:  tato

n=1:  tato
       (and tato only)

n=2:  tato
       (and tato only)
        (and tato (and tato only) only)

n=3:  tato
       (and tato only)
        (and tato (and tato only) only)
         (and tato (and tato only) (and tato (and tato only) only) only)
```

A short table of recursive expansions of the Lisp variable designated "tato"

The embryonic case occurs when longlist is nil. Then nothing happens, and so our answer should be nil.

The recursive case involves building a more complex answer out of a simpler one assumed to be given. We can fall back on our "(a b a)" example for this. We built the complex answer "(1 2 3 b 1 2 3)" out of the simpler answer "(b 1 2 3)" by appending "(1 2 3)" to it. On the other hand, we could consider "(b 1 2 3)" itself to be a complex answer built out of the simpler answer "(1 2 3)" by consing "b" onto it. Why does one involve appending and the other involve consing? Simply because the first case involves the atom "a", which *does* get replaced, whereas the second involves the atom "b", which does *not* get replaced. This observation enables us to attempt to write down a recursive definition of "replace", as follows:

```
(def replace
 (lambda (atm lyst longlist)
  (cond
   ((null longlist) nil)
   ((eq (car longlist) atm)
    (append
     lyst
     (replace atm lyst (cdr longlist))))
   (t
    (cons
     (car longlist)
     (replace atm lyst (cdr longlist)))))
  ))
```

As you can see, there is an embryonic case (where longlist equals nil) and then one recursive case featuring "append" and one recursive case featuring "cons". Now let us try out this definition on a new example:

```
—> (replace 'a '(1 2 3) '(a (a) b a))
(1 2 3 (a) b 1 2 3)
—>
```

Whoops! It *almost* worked, except that one of the occurrences of "a" was completely missed. This means that in our definition of "replace" we must have overlooked some eventuality. Indeed, if you go back, you will see that an unwarranted assumption slipped in right under our nose, namely that the elements of longlist are always atoms. We ignored the possibility that longlist might contain sublists. What to do in such a case? The answer: Do the replacement inside those sublists as well. And inside sublists of sublists too—and so on. Can you figure out a way to fix up the definition?

We have seen a recursion before where all parts on all levels of a structure needed to be explored; it was the function "atomcount" last month, in which we did a simultaneous recursion on both "car" and "cdr". The recursive line ran "(plus (atomcount (car s)) (atomcount (cdr s)))". Here it will be quite analogous. We shall have a recur-

sive line featuring *two* calls on "replace", one involving the car of longlist and one involving the cdr, instead of just one involving the cdr. This makes perfect sense when you think about it. Suppose you wanted to replace all the unicorns in Europe with porpuquines. One way to achieve this nefarious goal would be to split Europe into two pieces: Portugal (Europe's car) and the remainder (Europe's cdr). After replacing all the unicorns in Portugal with porpuquines, and also all the unicorns in the rest of Europe with porpuquines, you would finally recombine the two new pieces into a reunified Europe. (This is supposed to suggest a "cons" operation.) Of course, to carry out this dastardly operation on Portugal an analogous splitting and rejoining would have to take place—and so on. That suggests our recursive line will look like this:

```
(cons (replace 'unicorn
               '(porpuquine)
               (car geographical-unit))
      (replace 'unicorn
               '(porpuquine)
               (cdr geographical-unit)))
```

Or, more elegantly and more generally:

```
(cons (replace atm lyst (car longlist))
      (replace atm lyst (cdr longlist)))
```

This "cons" line will cover the case where longlist's car is nonatomic as well as the case where it is atomic but not equal to atm. In order to make this work we need to augment the embryonic case slightly: we shall say that when longlist is not a list but an atom, "replace" has no effect on longlist at all. Conveniently this subsumes the earlier "null" line, and so we can drop that one. If we put all of this together, we come up with a new, improved definition:

```
(def replace
 (lambda (atm lyst longlist)
  (cond
   ((atom longlist) longlist)
   ((eq (car longlist) atm)
    (append
     lyst
     (replace atm lyst (cdr longlist))))
   (t
    (cons
     (replace atm lyst (car longlist))
     (replace atm lyst (cdr longlist)))))
 ))
```

Now when we say "(tato-expansion 2)" to the Lisp genie, it will print out for us the list "(tato (and tato only) (and tato (and tato only) only))".

Well, well. Isn't this a magnificent accomplishment? If it seems less than magnificent, perhaps we can carry it a step further. A recursive acronym—one containing a letter standing for the acronym itself—can be amusing, but what about *mutually* recursive acronyms? This could mean, for instance, two acronyms each of which contains a letter standing for the other acronym. An example would be the pair of acronyms "NOODLES" and "LINGUINI", standing respectively for:

"NOODLES (oodles of delicious
    LINGUINI), elegantly served"

"luscious itty-bitty NOODLES got
    usually in Naples, Italy"

Notice, incidentally, that "NOODLES" is not only indirectly recursive but also directly so. There is nothing wrong with that.

In general the notion of mutual recursion means a system of arbitrarily many interwoven structures each of which is defined in terms of one or more members of the system (possibly including itself). If we are speaking of a family of mutually recursive acronyms, this means a collection of words in any one of which letters can stand for any word in the family.

I have to admit that this specific notion of mutually recursive acronyms is not particularly useful in any practical sense. It is nonetheless quite useful as a droll example of a very common abstract phenomenon. Who has not at some time mused about the inevitable circularity of dictionary definitions? Anyone can see that all words eventually are defined in terms of some fundamental set that is not further reducible but simply goes round and round endlessly. You can amuse yourself by looking up the definition of a common word in a dictionary and replacing the main words in it by *their* definitions. I once carried this process out for "love," defined as "a strong affection for or attachment or devotion to a person or persons," substituting for "strong," "affection," "attachment," "devotion" and "person" and coming up with this concoction: "A morally powerful mental state or tendency, having strength of character or will for, or affectionate regard, or loyalty, faithfulness or deep affection to, a human being or beings, especially as distinguished from a thing or lower animal."

Not being satisfied with that, I carried the entire process a step further. This was my result: "A set of circumstances or attributes characterizing a person or thing at a given time in, with, or by the conscious or unconscious together as a unit full of or having a specific ability or capacity in a manner relating to, dealing with, or capable of making the distinction between right and wrong in conduct, or an inclination to move or act in a particular direction or way, having the state or quality of being strong in moral strength, self-discipline, or fortitude, or the act or process of volition for, or consideration, attention, or concern full of fond or tender feeling for, or the quality, state, or instance of being faithful to, those persons or ideals that one is under obligation to defend or support, or the condition, quality or state of being worthy of trust, or a strongly felt fond or tender feeling to a creature or creatures of or characteristic of a person or persons, that lives or exists, or is assumed to do so, particularly as separated or marked off by differences from that which is conceived, spoken of, or referred to as existing as an individual entity, or from any living organism inferior or in rank, dignity, or authority, typically capable of moving about but not of making its own food by photosynthesis."

Isn't it romantic? It certainly makes "love" ever more mysterious. Stuart Chase, in his lucid classic on semantics, *The Tyranny of Words*, does a similar exercise for "mind" and shows its opacity equally well. Of course, concrete words as well as abstract ones get caught in this vortex of confusion. My favorite example is one I discovered while looking through a French dictionary many years ago. It defined the verb *clocher* ("to limp") as *marcher en boitant* (roughly "to walk while hobbling") and *boiter* ("to hobble") as *clocher en marchant* ("to limp while walking"). This eager learner of French was helped precious little by that particular pair of definitions.

But let us return to mutually recursive acronyms. I put quite a bit of effort into working out a family of them, and to my surprise they wound up dealing mostly (although by no means exclusively!) with Italian food. It all began when, inspired by "tato", I chose the similar word "tomato" and then decided to use its plural, coming up with this meaning for "tomatoes":

TOMATOES on MACARONI
    (and TOMATOES only),
    exquisitely SPICED

The capitalized words within the phrases are those that are also acronyms. Here is the rest of my mutually recursive family:

MACARONI:
    MACARONI and CHEESE
    (a REPAST of Naples, Italy)

REPAST:
    rather extraordinary PASTA
    and SAUCE, typical

CHEESE:
    cheddar, havarti, Emmentaler
    (particularly SHARP Emmentaler)

SHARP:
    strong, hearty and rather pungent

SPICED:

sweetly pickled in CHEESE
ENDIVE dressing

ENDIVE:
egg NOODLES, dipped
in vinegar eggnog

NOODLES:
NOODLES (oodles of delicious
LINGUINI), elegantly served

LINGUINI:
LAMB CHOPS
(including NOODLES),
got usually in northern Italy

PASTA:
PASTA and SAUCE (that's ALL!)

ALL!:
a luscious lunch

SAUCE:
SHAD and unusual COFFEE
(eccellente!)

SHAD:
SPAGHETTI, heated al dente

SPAGHETTI:
standard PASTA, always good, hot
particularly (twist, then ingest)

COFFEE:
choice of fine flavors,
particularly ESPRESSO

ESPRESSO:
excellent, strong, powerful, rich
ESPRESSO, suppressing sleep
outrageously

BASTA!:
belly all stuffed (tummy ache!)

LAMB CHOPS:
LASAGNE and meatballs,
casually heaped onto
PASTA SAUCE

LASAGNE:
LINGUINI and SAUCE
and GARLIC
(NOODLES everywhere!)

RHUBARB:
RAVIOLI, heated under butter
and RHUBARB (BASTA!)

RAVIOLI:
RIGATONI and vongole in oil,
lavishly introduced

RIGATONI:
rich Italian GNOCCHI
and TOMATOES
(or NOODLES instead)

GNOCCHI:
GARLIC NOODLES
over crisp CHEESE,
heated immediately

GARLIC:
green and red LASAGNE
in CHEESE

Any gourmet can see that little attempt has been made to have each term defined by its corresponding phrase; it is simply associated more or less arbitrarily with the phrase.

What happens if we begin to expand some word, say "pasta"? At first we get simply "PASTA and SAUCE (that's ALL!)". The next stage yields "PASTA and SAUCE (that's ALL!) and SHAD and unusual COFFEE (eccellente!) (that's a luscious lunch)". We could obviously go on expanding acronyms forever, or at least until we filled the universe to its very brim with mouth-watering descriptions of Italian food. But what if we were less ambitious and wanted merely to fill half a page or so with such a description? How might we find a way to halt this seemingly bottomless recursion in midcourse?

The key word here is "bottomless," and the answer it implies is: Put in a mechanism to allow the recursion to bottom out. The bottomlessness comes from the fact that at every stage every acronym is allowed to expand, that is, to spawn further acronyms. What if instead we kept tight control of the spawning process, being generous in the first few "generations" and gradually letting fewer and fewer acronyms spawn progeny as the generations got later? It would be similar to a redwood tree in a forest, which begins with a single "branch" (its trunk), and that branch spawns "progeny," namely the first generation of smaller branches, and they in turn spawn ever more progeny—but eventually a natural bottoming out comes as a consequence of the fact that teeny twigs simply cannot branch further. (Somehow in the course of evolution trees seem to have got their wires crossed, since for them bottoming out generally takes place at the top.)

If this process were completely regular, all redwood trees would look exactly alike and one could agree with former Governor Reagan's memorable dictum "If you've seen one redwood tree, you've seen them all." Unfortunately redwood trees (and maybe some other things) are trickier than Governor Reagan realized, and we have to learn to deal with a variety of things that go by the same name. The variety is caused by the introduction of randomness into the choices of whether to branch or not to branch, what angle to branch at, what size branch to grow and so on.

Similar remarks apply to the "trees" of mutually recursive acronyms. If in expanding "tomatoes" we always made exactly the same control decisions about which acronyms to expand when, there would be one and only one type of "rhubarb" expansion, so that here too it would make sense to say, "If you've seen one rhubarb, you've seen them all." But if we allow some randomness to enter the decision making about spawning, we can get many varieties of rhubarb, all bearing some telltale resemblance to one another but at a much more elusive level of perception.

How can we do it? The ideal concept to bring to bear here is that of the random-number generator, which serves as the computational equivalent of a coin flip or a dice throw. We shall let all decisions about whether or not to expand a given acronym depend on the outcome of such a virtual coin flip. At early stages of expansion we shall set things up so that the coin will be very likely to come up heads (do expand); at later stages it will be increasingly likely to come up tails (no expansion). The Lisp function "rand" will be employed for this. It takes no arguments, and each time it is called it returns a new real number somewhere between 0 and 1, unpredictably. (This is an exaggeration. It is actually 100 percent predictable if you know how it is computed, but since the algorithm is rather obscure, for most purposes and for most observers the behavior of the function will be so erratic that it counts as being totally random. The story of random-number generation is quite a fascinating one and would be an article in itself.)

If we want an event to happen with a probability of 60 percent, first we ask rand for a value. If the value turns out to be .6 or below, we go ahead, otherwise we do not. Since over a long period of time rand sprays its outputs uniformly over the interval between 0 and 1, we shall indeed get the go-ahead 60 percent of the time.

So much for random decisions. How do we get an acronym to expand when it is told to? That is not too hard. Suppose we let each acronym be a Lisp function, as in the following example:

```
(def tomatoes
 (lambda ( )
  '(tomatoes on macaroni
    (and tomatoes only)
    exquisitely spiced))))
```

The function "tomatoes" takes no arguments and simply returns the list of words it expands into. Nothing could be simpler.

Now suppose we have a variable called "acronym" whose value is some particular acronym—but we do not know which one. How could we get that acronym to expand? The way we have set it up the acronym must act as a function call. In order for any atom to invoke a function it must be the car of a list, as in the examples "(plus 2 2)", "(rand)" and "(rhubarb)". If we were to write "(acronym)", the literal atom "acronym" would be taken by the Lisp genie

From 0 to 55,134 feet in exactly 1 hour 9 minutes and 42 seconds.

Julian Nott makes final pre-flight checks.

Nothing inspires a world record holder more than the capture of his world record.

In 1979, it happened to Julian Nott.

A competitor soared to 52,000 feet in a hot air balloon, exceeding Nott's 1974 record by more than 6,000 feet.

But Julian Nott refused to accept defeat.

Working with a team of dedicated experts, he spent 13 months meticulously planning to regain the record.

If he was to succeed, he would have to pilot his balloon into the top ten percent of the atmosphere. A barometric desert where freezing temperatures, lack of oxygen and virtually non-existent air pressure pose a constant threat to both man and machine.

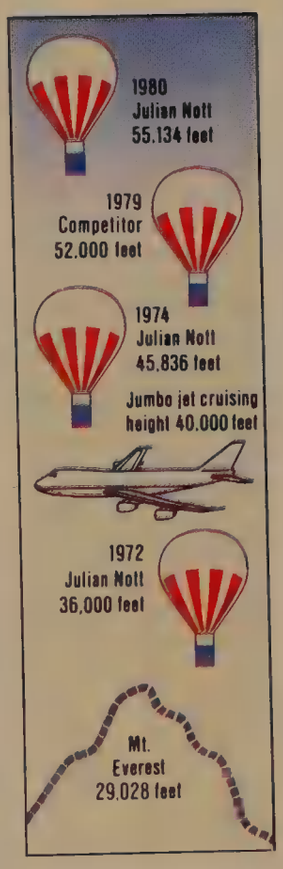
Every piece of equipment for the ascent was chosen with agonizing attention to detail.

The most advanced technology and materials were incorporated.

For example, the pressurized cabin was constructed of the very latest composite plastic, effecting a weight saving of nearly 80 percent over conventional materials.

One vital piece of Julian Nott's equipment, however, remained unchanged from his 1974 ascent.

His watch. A Rolex Oyster. "It came under the same close scrutiny as everything else," remarked Julian, "but, personal preferences aside, its inclusion was never really in doubt."



The record since 1972.

In 1980, in the early hours of a late October morning, in Denver, Julian Nott and his hot air balloon *Innovation* began the record-breaking attempt.

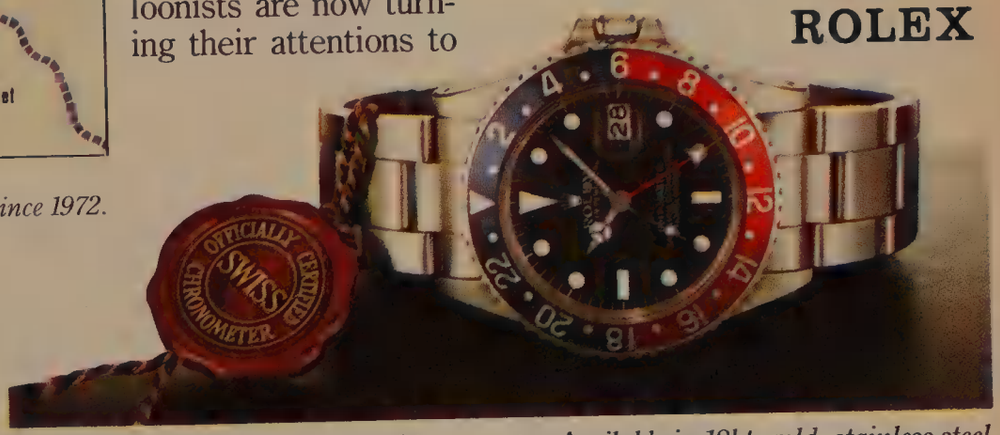
Using his Rolex Oyster to enable him to always match the computer-designated optimum rate of climb, he rose to a height of 55,134 feet. A new world record.

The world's top balloonists are now turning their attentions to the circumnavigation of the earth.

Julian Nott is one of them.

"Of course, choice of equipment will be crucial for this trip too," he says. "But I'll be very surprised if the first watch to go around the world by balloon isn't a Rolex." He smiles. "Even if it's not mine."

ROLEX



The Rolex GMT-Master Oyster Chronometer. Available in 18kt. gold, stainless steel or steel and gold combination. All with matching bracelets.

Write for brochure. Rolex Watch, U.S.A., Inc., Dept. 312 Rolex Building, 665 Fifth Avenue, New York, N.Y. 10022-5383.

as a function name. That, however, would be a misunderstanding. It is certainly not the atom "acronym" that we want to make serve as a function name, but its *value,* be it "macaroni", "cheese" or what have you.

In order to accomplish this we resort to a little trick. If the value of the atom "acronym" is "rhubarb" and if I write "(list acronym)", the value the genie will return to me will be the list "(rhubarb)". The genie, however, will simply see this as an inert piece of Lispstuff rather than as a little command I should like to have executed. It cannot read my mind. How do I get it to perform the desired operation? The answer is that I remember the function called "eval", which makes the genie look on a given data structure as a wish to be executed. In this case I need merely say "(eval (list acronym))" and I shall get the list "(ravioli, heated under butter and rhubarb (basta!))". And if the acronym had had a different value, the genie would have handed me a different list.

We now have just about enough ideas to build a function capable of expanding mutually recursive acronyms into long but finite phrases whose sizes and structures are determined by many "flips" of the "rand" coin. Instead of stepping you through the construction of this function I shall simply display it and let you peruse it. It is modeled very closely on the earlier function "replace":

```
(def expand
 (lambda (phrase probability)
  (cond
   ((atom phrase) phrase)
   ((is-acronym (car phrase))
    (cond
     ((lessp (rand) probability)
      (append
       (expand (eval (list (car phrase)))
               (lower probability))
       (expand (cdr phrase) probability))))
     (t
      (cons (car phrase)
            (expand (cdr phrase)
                    probability))))))
  (t
    (cons
     (expand (car phrase)
             (lower probability))
     (expand (cdr phrase) probability))))
))
```

Note that "expand" has two parameters. One parameter represents the phrase to be expanded, the other represents the probability of expanding any acronyms that are top-level members of the given phrase. (Thus the value of the atom "probability" will always be a real number between 0 and 1.) As in the redwood-tree example, the expansion probability should decrease as the calls get increasingly recursive. This is why lines that call for the expansion of (car phrase) do so with a *lowered* probability. To be exact, we can define the function "lower" as follows:

```
(def lower (lambda (x) (times x .8)))
```

Thus each time an acronym expands, its progeny are only .8 times as likely to expand as it was. This means acronyms nested deep enough have a vanishingly small probability of spawning further progeny. You could use any reducing factor; there is nothing sacred about .8 except that it seems to yield good results for me.

The only remaining undescribed function inside the definition above is "is-acronym". Its name is pretty self-explanatory. First the function tests to see if its argument is an atom; if it is not, the function returns nil. If the argument *is* an atom, the function goes on to see whether that atom has a function definition, in particular a definition with the form of an acronym. If the atom has such a definition, is-acronym returns the value t; otherwise it returns nil. Precisely how this is done depends on your specific variety of Lisp, which is why I have not shown it explicitly. In Franz Lisp only one line is needed.

You may have noticed that there are two cond clauses in close proximity that begin with "t". How come one "otherwise" follows so closely on the heels of another one? Actually they belong to different conds, one nested inside the other. The first "t" (belonging to the inner cond) applies to a case where we know we are dealing with an acronym but where our random coin, instead of coming down heads, has come down tails (which amounts to a decision not to expand); the second "t" (belonging to the outer cond) applies to a case where we have discovered we are simply not dealing with an acronym at all.

The inner logic of "expand", when it is scrutinized carefully, makes perfect sense. On the other hand, no matter how carefully you scrutinize it, the output produced by "expand" using this "famiglia" of acronyms remains quite silly. Here is an example:

(rich italian green and red linguini and shad and unusual choice of fine flavors, particularly excellent, strong, powerful, rich espresso, suppressing sleep outrageously (eccellente!) and green and red lasagne in cheese (noodles everywhere!) in cheddar, havarti, Emmentaler (particularly sharp Emmentaler) noodles (oodles of delicious linguini), elegantly served (oodles of delicious linguini), elegantly served (oodles of delicious linguini and sauce and garlic (noodles (oodles of delicious linguini), elegantly served everywhere!) and meatballs, casually heaped onto pasta and sauce (that's all!) and sauce (that's a luscious lunch) sauce (including noodles (oodles of delicious linguini), elegantly served), got usually in northern Italy), elegantly served over crisp cheese, heated immediately and tomatoes on macaroni and cheese (a repast of Naples, Italy) (and tomatoes only), exquisitely sweetly pickled in cheese endive dressing (or noodles instead) and vongole in oil, lavishly introduced, heated under butter and rich Italian gnocchi and tomatoes (or noodles instead) and vongole in oil, lavishly introduced, heated under butter and rigatoni and vongole in oil, lavishly introduced, heated under butter and ravioli, heated under butter and rich Italian garlic noodles over crisp cheese, heated immediately and tomatoes (or noodles instead) and vongole in oil, lavishly introduced, heated under butter and ravioli, heated under butter and rhubarb (basta!) (basta!) (basta!) (basta!) (belly all stuffed (tummy ache!)) (basta!))

Can you figure out which acronym this gastronomical monstrosity grew out of? As you can see, Lisp is hardly the computer language to learn if you want to lose weight. One final example of the glories of recursive spaghetti is given herewith, expanded from a different starting point:

(macaroni and cheese (a rather extraordinary pasta and sauce, typical of Naples, Italy) and cheddar, havarti, Emmentaler (particularly sharp Emmentaler) (a rather extraordinary pasta and shad and unusual coffee (eccellente!) (that's a luscious lunch) and sauce (that's all!) and shad and unusual choice of fine flavors, particularly espresso (eccellente!) (that's all!) and sauce (that's a luscious lunch) and spaghetti, heated al dente and unusual choice of fine flavors, particularly excellent, strong, powerful, rich espresso, suppressing sleep outrageously (eccellente!), typical of Naples, Italy))

The "expand" function exploits one of the most powerful features of Lisp. That is the ability of a Lisp program to take data structures it has created and treat them as pieces of code (that is, give them to the Lisp genie as commands). Here it was done in a most rudimentary way. An atom was wrapped in parentheses and the resulting minuscule list was then evaluated, or "eval'd", as Lispers' jargon has it. The work involved in manufacturing the data structure was next to nothing in this instance, but in other instances elaborate pieces of structure can be "consed up", then handed to the Lisp genie for "eval'ing". Such pieces of code might be new function definitions or any number of other things. The main idea is that in Lisp one has the ability to "elevate" an inert, information-containing data structure to the level of an "animate agent," where it becomes a manipulator of inert structures itself. This program-data cycle, or loop, can con-

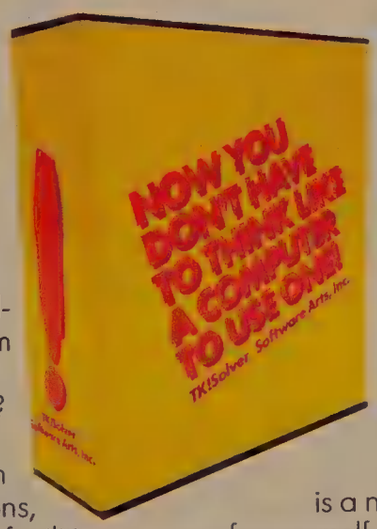
CRUNCH!

The inventor of VisiCalc®, the world's leading software product, introduces TK!Solver™—the revolutionary way to crunch problems on a personal computer.

With VisiCalc, Software Arts™ created an entirely new application for personal computers—the electronic spreadsheet.

Now Software Arts has developed TK!Solver, another entirely new and different application for personal computers—the problem solver. Business and financial problems. Scientific and engineering problems. Architectural and building design problems. Any kind of problem you can dream up, in any area.

What's so revolutionary about that? Simply this: the power of the TK!Solver program comes from the ease with which you can set up problems, vary assumptions, find solutions, and display results. All the facilities needed to solve the problems—simple or complex—are built-in, and need not be developed. Once you have stated your problem, simply enter the known values, then solve the problem with a single keystroke.



For even easier problem solving use a TK!SolverPack™ in which models for solving common problems in such areas as finance, engineering and education have already been formulated by professionals in these fields.

There is no software product in existence that solves problems as quickly, flexibly, and easily as TK!Solver—The Problem Cruncher™. In fact, no other software product even works the way TK!Solver does. It is a necessity for personal computer owners. See for yourself. Visit your nearest computer store for a demonstration and find out how much you can do with a personal computer when it thinks as flexibly as you do.

Software Arts™

27 Mica Lane, Wellesley, MA 02181

TK!Solver, TK!SolverPack and TK!Solver—The Problem Cruncher are trademarks of Software Arts, Inc. Software Arts is a trademark of Software Arts, Inc. and Software Arts Products Corp. VisiCalc is a trademark of VisiCorp. Copyright © 1983. Software Arts, Inc. All rights reserved.

80 PROOF · CALVERT DISTILLERS CO., NYC

After the day's hassle and hustle, try a soothing change of pace.

John Jameson
Imported Irish Whiskey

tinue on and on, with structures reaching out, twisting back and indirectly modifying themselves or structures related to them.

Certain types of inert, or passive, information-containing data structures are sometimes referred to as "declarative knowledge"—"declarative" because they often have a form abstractly resembling that of a declarative sentence, and "knowledge" because they encode facts about the world, accessible by looking in an index in somewhat the way "book-learned" facts are accessible to a human being. In contrast, animate, or active, pieces of code are referred to as "procedural knowledge"—"procedural" because they define sequences of actions ("procedures") that actually manipulate data structures, and "knowledge" because they can be viewed as embodying the program's set of skills, something like a human being's unconscious skills that were once learned through long rote drill sessions. Sometimes these contrasting knowledge types are referred to as "knowledge that" and "knowledge how."

This distinction should remind biologists of the distinction between genes, which are relatively inert structures inside the cell, and enzymes, which are anything but inert. Enzymes are the animate agents of the cell; they transform and manipulate all the inert structures in indescribably sophisticated ways. Moreover, Lisp's loop of program and data should remind biologists of the way genes dictate the form of enzymes and enzymes manipulate genes (among other things). Thus Lisp's procedural-declarative program-data loop provides a primitive but very useful and tangible example of one of the most fundamental patterns at the base of life: the ability of passive structures to control their own destiny, by creating and regulating active structures whose form they dictate.

We have been talking all along about the Lisp genie as a mysterious given agent, without asking where it is to be found or what makes it work. It turns out that one of Lisp's most exciting properties is that one can describe with ease the Lisp genie's complete nature in Lisp itself. In other words, the Lisp interpreter can be easily written down in Lisp. Of course, if there is no Lisp interpreter to run *that* Lisp interpreter, it might seem like an absurd and pointless exercise, a bit like having a description in flowery English telling foreigners how best to learn English. It is not, however, as silly as that makes it sound.

In the first place, if you know enough English, you can "bootstrap" your way further into English; there is a point beyond which explanations written in English about English are indeed quite useful. What is more, that point is not too far beyond the beginning level. Therefore all you need to acquire first, and autonomously, is a "kernel"; then you can begin to lift yourself by your own bootstraps. For children it is an exciting thing when, as they read, they begin to learn new phrases all by themselves, simply by encountering them several times in succession. Their vocabulary begins to grow by leaps and bounds. So it is once there is a Lisp kernel in a system; the rest of the Lisp interpreter can be written in Lisp and usually is.

The fact that one can easily write the Lisp interpreter in Lisp is no mere fluke of some peculiarly introverted fact about Lisp. The reason it is easy is that, because of its program-data loop, Lisp lends itself to writing interpreters for all kinds of computer languages. This means Lisp can serve as a basis on which one can build other languages.

To put it more vividly, suppose you have designed on paper a new language called "Flumsy." If you really know how Flumsy should work, it should not be too hard for you to write an interpreter for it in Lisp. Once your Flumsy interpreter is implemented, it becomes in essence a new genie to which you can give wishes in Flumsy and that will in turn communicate those wishes to the Lisp genie in Lisp. Of course, all the mechanisms allowing the Flumsy "meta-genie" to talk with the Lisp genie are themselves being carried out by the Lisp genie. Is this a mere facade? Is talking Flumsy really just a way of talking Lisp in disguise?

Well, when the U.S. arms negotiators talk with their Russian counterparts through an interpreter, are they really just speaking Russian in disguise? Or is the crux of the matter whether the interpreter's native language was English or Russian, on which the other language was built as a second one? And suppose you find out that actually the interpreter's native language was Lithuanian, that she learned English only as an adolescent and then learned Russian by taking high school classes taught in English? Will you then think that when she speaks Russian, she is actually speaking English in disguise, or worse, that she is actually speaking Lithuanian, doubly disguised?

Analogously, you might find that the Lisp interpreter is actually written in Pascal or some other computer language. And then someone could strip off the Pascal facade as well, revealing to you that all instructions are really being executed in *machine language,* so that you are fooling yourself completely if you think the machine is talking Flumsy, Lisp, Pascal or any other higher-level computer language!

When one interpreter runs on top of another, there is always the question of what level one chooses not to look below. I personally seldom think about what underlies the Lisp interpreter, so that when I am dealing with the Lisp

Phone Extortion

Let's put the screws to our phone suppliers on a terrific long range cordless phone with a novel new consumer buying tool called GREED!

It's a war. The major phone suppliers are each desperately trying to lock up big orders. And now, we've devised a plan to take advantage of their greed.

Together we can apply enough pressure to get you a super high quality cordless phone with a range up to 700' for just $79. And, hopefully we'll make a good profit for ourselves. But, there's a problem.

TIGHTEN THE SCREWS

Before I tell you all about the great features and sound of the phone, let me explain just how we plan to deliver these phones for such an admittedly foolhardy price. You see, it's called extortion (legal definition please) and it works like this.

There are two large phone companies that are trying to get DAK's phone business. To be perfectly honest, the quality, guarantee and sound are so close that we couldn't decide between the two.

Even their prices are close. So here's what we did. We got quotes from both of them for 5,000, then for 50,000 phones.

And, here's what we decided. As you place your orders, we'll offer each of the two companies a check (cash in advance) for the number of phones we need.

The company with the **lowest price of the day** will get that day's order. You see, your orders give us incredible leverage.

Usually we buy just one phone from one supplier. Unfortunately once they have a contract with us, they have no incentive to lower our price. And, at our costs today, we really can't come out at only $79.

SAFE INVESTMENT

Don't worry about your $79. DAK is a large company. If this plan to sell 50,000 phones fizzles out, we'll still be OK. Plus, DAK doesn't even charge your credit card until **after** we ship.

And most important. Each phone will come to you in its factory sealed carton, and will be backed by the manufacturer's standard limited warranty.

Finally, you won't be getting a cheap stripped down phone. We aren't dumb. We picked the phones we wanted **before** we got the quotes. So, nobody can make any changes to come out lowest.

And frankly we'll get 50,000 of you new customers (we have 225,000 now) to send our 68 page electronics catalog to. So, even if it does fizzle, we're sure to get another chance at you later.

A GREAT LONG RANGE PHONE

So, now let's see if this is really going to be a stripped down low quality phone. Remember, each and every feature I describe appears on both phones. And, if for some emotional reason you want to specify which phone you want, no problem, I've got a way for you to do that too.

Imagine walking from room to room in your home talking on the phone. You'll go out and get the mail, walk over to a neighbors or work in the yard or the garage.

It's a superb personal security device. You can instantly call a doctor or police from anywhere you happen to be.

And, if you are needed at home, no problem. There's a page button on the base that beeps you at the handset which then becomes a wireless intercom.

Cordless phones are a great combination of posh luxury and life saving utility.

NO INSTALLATION

If you can plug in a lamp you can install your new cordless phone. Just plug in the modular phone plug and the AC line cord, then start talking. The phone uses universal pulse dialing compatible with both rotary dial and Touch Tone® phone lines.

LOADED WITH FEATURES

Look at these features.You'll have last number redial to redial busy numbers. And mute lets you electronically put your hand over the mouthpiece for privacy.

There's both a standby/talk switch and a power off switch so that you don't have to return the handset to the base when you aren't using it.

And when you do put the handset in the base, its heavy duty NiCad batteries recharge automatically, and the auto secure protects your base station.

If you're adjusting a TV antenna, the person at the base can talk to you using the intercom feature and any of your regular house phones. So, you'll have a cordless intercom, with a range up to 700 ft too.

The two companies that we are trying to 'extort' our low prices from are Unitech, and U.S. Tron. Even their names sound alike. Both phones are backed by a manufacturer's standard limited warranty.

TRY TOTAL PHONE FREEDOM RISK FREE

Experience the thrill of total phone freedom as you roam throughout your home, yard or even a neighbor's house. You'll never have to 'run for the phone' again.

Try a cordless phone in your own home risk free. If you don't find the luxury of talking from wherever you are 100% enthralling, just return the phone to DAK in its original box within 30 days for a refund.

To order a long range full featured cordless phone risk free with your credit card, call toll free or send your check for the 'extorted' price of just **$79** plus $3.50 for postage and handling. Order Number 9615. CA res add 6% sales tax.

If you have a particular love for either phone, you can have your choice for just $89 plus $3.50 for postage and handling. To order the U.S. Tron, use Order Number 9616. To order the Unitech, use Order Number 9617. So, you'll pay only $79 if the choice is ours, or $89 if it's yours.

DAK
INDUSTRIES INCORPORATED
TOLL-FREE 1-800-423-2636
If busy, after hours, on weekends or in CA
CALL TOLL-FREE . . . 1-800-228-1234
10845 Vanowen St., N. Hollywood CA 91605

INVEST YOURSELF

A windmill to pump water for "salt farming" in India. More efficient woodburning stoves for the Sahel. Photovoltaic irrigation pumps for the Somali refugee camps.

All these are solutions to technical problems in developing countries. Devising such solutions is no simple task. To apply the most advanced results of modern science to the problems of developing areas in a form that can be adopted by the people requires the skills of the best scientists, engineers, farmers, businessmen—people whose jobs may involve creating solid state systems or farming 1000 acres, but who can also design a solar still appropriate to Mauritania or an acacia-fueled methane digester for Nicaragua.

Such are the professionals who volunteer their spare time to Volunteers in Technical Assistance (VITA), a 20 year old private, non-profit organization dedicated to helping solve development problems for people world-wide.

Four thousand VITA Volunteers from 82 countries donate their expertise and time to respond to the over 2500 inquiries received annually. Volunteers also review technical documents, assist in writing VITA's publications and bulletins, serve on technical panels, and undertake short-term consultancies.

Past volunteer responses have resulted in new designs for solar hot water heaters and grain dryers, low-cost housing, the windmill shown above and many others. Join us in the challenge of developing even more innovative technologies for the future.

VITA Putting Resources to Work for People

3706 Rhode Island Ave, Mt. Rainier, Maryland 20712 USA

system, I feel as if I am talking with "someone" whose "native language" is Lisp. Similarly, when I am dealing with people, I seldom think about what their brains are composed of; I do not reduce them mentally to piles of patterned neuron firings. It is natural to my perceptual system to recognize them at a certain level and not to look below that level.

If someone were to write a program that could deal in Chinese with simple questions and answers about restaurant visits, and if that program were written in the hypothetical language "SEARLE" (for "Simulated East-Asian Restaurant-Lingo Expert"), I could choose to view the system either as genuinely speaking Chinese (assuming it gave a creditable and not too slow performance) or as genuinely speaking SEARLE. I can shift my point of view at will. The one I adopt is governed mostly by pragmatic factors, such as which subject area I am currently more interested in (Chinese restaurants or how grammars work), how closely matched the given level's speed is to that of my own brain and, not least, whether I happen to be more fluent in Chinese or in SEARLE. If to me Chinese is a mere bunch of squiggles and squoggles, I may opt for the SEARLE viewpoint; if on the other hand SEARLE is a mere bunch of confusing technical expressions, I shall probably opt for the Chinese viewpoint. And if I find out that the SEARLE system is in turn implemented on top of a Lisp system, then I have yet a third point of view to choose. And so on.

With interpreters stacked on interpreters, however, things rapidly become very inefficient. It is like running a motor on power coming from a series of electric generators, each generator being run on power coming from the preceding one: a good deal of power is lost at each stage. With generators there is usually no need for a long cascade, but with interpreters it is often the only way to go. If there is no machine whose machine language is Lisp, then you build a Lisp interpreter for whatever machine you have available and run Lisp that way. And Flumsy or SEARLE, if you want to have it at your disposal, is then built on top of this "virtual Lisp machine." Such indirectness can be annoyingly inefficient, causing your new virtual Flumsy machine or virtual SEARLE machine to run dozens of times slower than you would like.

There have been important hardware developments in the past several years, and now machines are available that are based on Lisp at the hardware level. This means that they "speak" Lisp in a somewhat deeper sense—let us say "more fluently"—than virtual Lisp machines. It also means that when you are working on such a machine, you are "swimming" in a Lisp environment. A Lisp environment goes considerably beyond what I have described so far, because it is more than just a language for writing programs. It includes an editing program, with which one can create and modify one's programs (and text as well), a debugging program, with which one can easily localize one's errors and correct them, and many other features, all designed to be compatible with one another and with an overarching "Lisp philosophy."

Such machines, although they are still expensive and somewhat developmental, are rapidly becoming cheaper and more available. They are put out by various new companies such as LMI (LISP Machine, Inc.), Symbolics, Inc., both of Cambridge, Mass., and older companies such as Xerox. Lisp is also available on most personal computers; you need only look at any issue of the current microcomputer magazines to find advertisements for it.

Why, in conclusion, is Lisp popular in artificial intelligence? There is no single answer, or even a simple answer. Would it be valid to say Lisp is "the language of thought"? Certainly not. AI people might have said so 10 or 20 years ago, but few would be so bold today. One reason Lisp has remained so popular is, as I wrote in my first Lisp column, that Lisp is crisp. Indeed, Lisp can be so elegant that a good Lisp function can please a Lisper as a poem pleases a lover of poetry. My colleague Dan Friedman is in fact compiling a book titled *Lisp Poems,* and he urges interested readers to send him their most elegant efforts. His address is Computer Science Department, Indiana University, Bloomington, Ind. 47405.

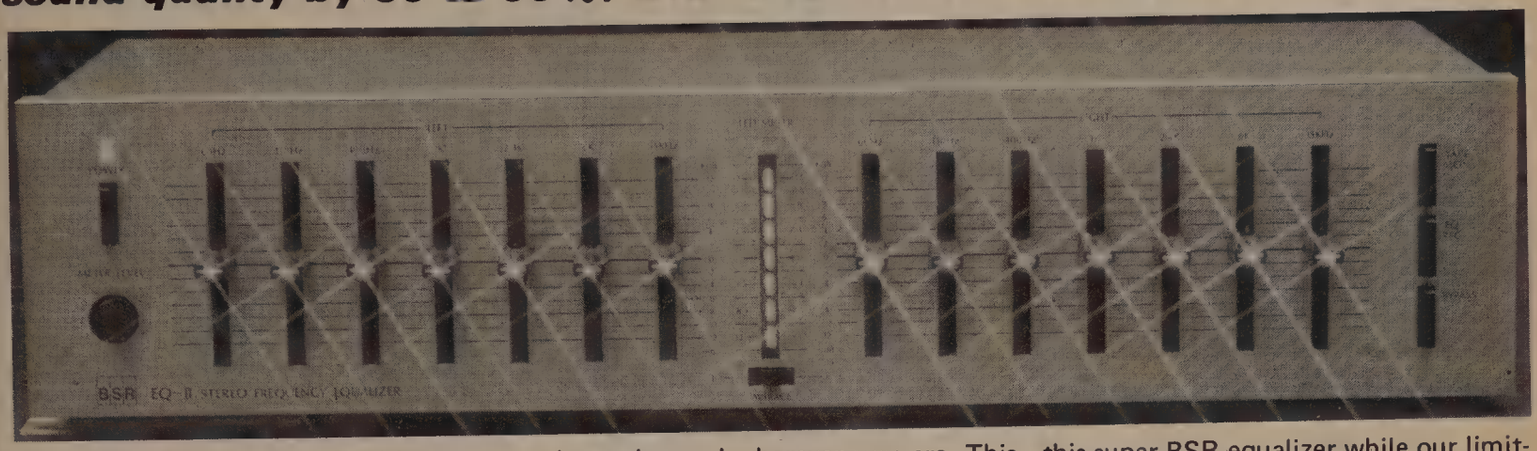
What properties of Lisp, then, have made it central to AI research? I would say the following ones are nearly exhaustive:

1. Lisp is crisp.
2. Lisp is interactive.
3. Lisp is centered on the idea of lists and their manipulation, and lists are extremely flexible, fluid data structures.
4. Lisp code, having the same form as Lisp data, can easily be manufactured in Lisp and run.
5. Interpreters for new languages can easily be built and experimented with in Lisp.
6. "Swimming" in a Lisplike environment feels natural for many people.
7. Lisp is permeated by the "recursive spirit."

Perhaps it is this last rather intangible statement that gets at it the best. For some reason many people in artificial intelligence seem to have a deep sense that recursivity, in *some* form or other, is connected with the "trick" of intelligence. It is a hunch, an intuition, a vaguely felt and slightly mystical belief, and one that I certainly share, but whether it will pay off in the long run remains to be seen.

$79 CLOSE-OUT
Sound Detonator

Make your stereo system's sound explode with life, improving the sound quality by 30 to 50%., with this limited $79 close-out from BSR.



It's like night and day. Crashing cymbals, the depth of a string bass, more trumpets or more voice. Now they'll all come bursting forth from your stereo system at your command.

You'll make your music so vibrant that it will virtually knock your socks off when you use this professional quality stereo Sound Detonator equalizer.

It has a frequency response from 5hz to 100,000hz ±1db. And, it's made and backed by a 2 year limited warranty by BSR, the ADC equalizer people. Our $79 close-out price is just a fraction of its true retail value.

CAN YOUR STEREO SOUND BETTER?

Incredibly better. Equalizers are very different from conventional bass and treble controls.

Bass controls turn up the entire low end as well as the low mid-range making the sound muddy and heavy. With an equalizer, you simply pick the exact frequencies you want to enhance.

You can boost the low bass at 60hz, and/or 150hz, and the mid-bass at 400 hz to animate specific areas of the musical spectrum.

And, best of all when you boost the part of the bass you like, you don't disturb the mid-range frequencies and make you favorite singer sound like he has a sore throat.

The high frequencies really determine the clarity and brilliance of your music. You can boost the high mid-range at 2400hz, or the high end at 6000hz and 15,000hz. So, you can bring crashing cymbals to life at 15,000hz while at the same time you cut tape hiss or annoying record scratches at 6000hz.

You can also boost or cut specific midrange frequency areas to add or subtract vocal, trumpets, guitars or whatever instrument ranges you prefer.

THERE'S MORE

You can push a button and transfer all the equalization power to the input of your tape deck. So, if you have a cassette deck in your car, or a personal stereo that you wear, now you can pre-equalize your cassettes as you record them with no cables to switch.

Now you can get all the dramatically enhanced sound wherever you are. This is an especially great feature for bass starved portables and high-end starved car stereos to make them come alive.

EASY HOOK UP

Use your tape monitor circuit, but don't lose it. Just plug the equalizer into the tape 'in' and 'out' jacks on your receiver. We even supply the cables.

As you listen to your records, FM stereo or 'aux', any time you push the tape monitor switch on your receiver you'll hear your music jump to life.

For your tape deck, simply plug it into the tape 'in' and 'out' jacks on the equalizer exactly as it was plugged into your receiver.

The output from your receiver is always fed direct to your tape deck for recording and with the touch of a button, you can choose to send equalized or nonequalized signal to your recorder.

When you want to listen to your tape deck, just press tape monitor on the equalizer and your tape deck will work **exactly** as it did before. Except, that now you can choose to listen to it with or without enhancing equalization.

You won't be listening to any distortion or hum. The Sound Detonator has a 95db signal to noise ratio and total harmonic distortion of only 0.018%.

Once you've set your equalizer controls, switch it in and out of the system. Then you'll hear such an explosive improvement in sound you'll think you've just added thousands of dollars of new speakers and equipment.

No one would believe that a $159 component let alone one at our $79 close-out price could do so much.

WHY A CLOSE-OUT

BSR is a very large company. Somebody decided to market equalizers under both their ADC subsidiary and the BSR names. Well, we never thought it was a very practical idea. And, now they seem to agree.

From now on you'll only see ADC equalizers. But, because they didn't know what to do with these that were labeled BSR, we got them for a song.

So, you can go to any HiFi store and buy an ADC equalizer made by the parent company BSR, or you can get this super BSR equalizer while our limited supply lasts for only $79.

Oh yes, if you want to know more about BSR, they also own DBX, the noise reduction company and if you're familiar with the X10 remote control system for your home, that's BSR too.

THE FINAL FACTS

There are 14 slide controls each with a bright LED to clearly show its position. Each control will add or subtract up to 12db. (That's a 24db range!)

There are separate sound detonation slide controls for each channel at 60hz, 150hz, 400hz, 1000hz, 2400hz, 6000hz, and 15,000hz.

There's an LED VU meter to show the relative channel output levels of the left, right and average of both. Plus there's a meter level control. It's 16-9/16" wide, 7-1/2" deep, and 3-9/16" tall.

PUT LIFE INTO YOUR MUSIC RISK FREE

Prepare for a shock the first time you switch in this equalizer. Instruments you never knew were in your music will emerge and bring a lifelike sound that will envelop you and revolutionize your concept of your home stereo.

Hook this BSR into your system and really give it a workout. If you aren't 100% satisfied for any reason, simply return it to DAK within 30 days in its original box for a courteous refund.

To order your Sound Detonator BSR EQ-2 Stereo Frequency Equalizer risk free with your credit card, simply call the DAK toll free hotline, or send your check for only **$79** plus ~~$6~~ for postage and handling. Order Number 9420. CA residents please add 6% sales tax.

Wake up the sound in your stereo. Your sound will explode with life as you detonate each frequency band with new musical life.

DAK
INDUSTRIES INCORPORATED
TOLL-FREE 1-800-423-2636
If busy, after hours, on weekends or in CA
CALL TOLL-FREE . . . 1-800-228-1234
10645 Vanowen St., N. Hollywood CA 91605

BOOKS

Crystals at high pressure, salt hunger, our star the sun, blueprints, the piano

by Philip Morrison

COMPARATIVE CRYSTAL CHEMISTRY: TEMPERATURE, PRESSURE, COMPOSITION AND THE VARIATION OF CRYSTAL STRUCTURE, by Robert M. Hazen and Larry W. Finger. John Wiley & Sons, Inc. ($43.95). Their eyes on the solid earth, mineralogists must pursue the behavior of mineral crystals under high pressure. A hundred kilobars (one bar is an atmosphere of load) takes the sample to an equivalent depth of 300 kilometers, within the earth's mantle, far beyond the dreams of the drillers. In this remarkable summary of the work of a decade or so, two deft and ingenious experimenters at the Carnegie Institution of Washington let us into a body of techniques all but magical in their simplicity and power. The second half of their book is a wide synopsis of the data obtained. The general reader will perhaps enjoy most the step-by-step explications of how it is all done.

The work of the high-pressure pioneers, like Percy Bridgman between the wars, was clever, heroic, clumsy. The big hydraulic press, kept behind a thick wall in case of catastrophic failure, strained against a complex fitting deep within which some sample was under stress. The behavior of the sample was monitored by a wild variety of obscure effects that could somehow be teased to signal out of the hidden region along a wire. Those measurements did not often reach the 100-kilobar range. Nowadays the work appears effortless, carried out under these crushing pressures in palm-sized apparatus loaded by tightening up by hand with a Crescent wrench!

The preconditions for this dazzling development are clear. The fundamental physical understanding is that a single crystal a tenth the bulk of a pinhead has enough atoms to display with precision the X-ray-diffraction pattern that reveals its atomic structure and spacing. The chief technical foundation is the long development of X-ray diffraction, so that nowadays automatic, sensitive photon-counting instruments provide speedy printout of structural information at high precision in hours, where older methods would have taken months of hand setting and hand computation. Finally, the strong bonds of the lightest atoms, realized in diamond and in boron carbide, supply us with materials that are the hardest, yet are entirely transparent to the X rays used to probe most important minerals.

To make a high-pressure, high-temperature cell able to hold a crystal in the analyzing beam it is enough to begin with a pair of quarter-inch slices cut out of hardened steel rod stock an inch and a half in diameter. Three half-inch cap screws serve to load the sandwich. Fitted with care into the steel disks is a multiple filling, first two boron carbide disks, then a refractory ring with heater wires. The tiny sample, along with two or three calibrating crystals, lies in the pressurized core. That cell is a one-third-millimeter hole in a foil of nickel alloy of about the same thickness. The ceiling and floor of the cell are two carefully prepared diamond anvils, borne between the boron carbide disks.

A methanol-based fluid fills the cell. Turn the cap screws in sequence a few degrees at a time, switch on the heater control circuitry and you can place in the path of a fine diffractometer an X-ray-transparent sample at the conditions of the earth's mantle. The diamond anvils, the heart of the device, must be well made, but they are not large enough for a modest engagement ring. The force they transmit is a few hundred pounds; the stored energy released when they fail is no larger than that of a laboratory notebook falling to the floor. The heroics have been supplanted by subtlety.

The technique is not exhausted. Hazen and Finger, the authors of this book, are working with crystallized gases, for example argon and neon, to replace the methanol pressure fluid. That change promises a pressure range a few times higher, at least at cryogenic temperatures. A tiny diamond-anvil system of somewhat more complicated mechanical design has reached the highest static pressures ever attained: close to two megabars.

The authors have made their book a how-to-do-it one. It has the flavor of old days at the laboratory bench, with its careful reminders of the right use for Vaseline, Kleenex, acetone, wax and cements, and close microscopic inspection as the marvelous structure is being made up out of its apt and precious parts. The second half of the book discusses results. A few dozen researches, often on a related family of inorganic compounds, are reported in summary, with tables and graphs. The authors have included all the high-temperature, high-pressure X-ray structure data they knew of up to mid-1981.

The results are satisfying: the unit polyhedrons that form many of the important oxides and silicates in this study allow rough prediction of the variations of structures with ambient conditions and with systematic chemical composition change. That justifies the title of the volume. Phase changes, complex and important as they are, are not yet so easy to tame, but continuous changes in structure suggest the rise of misfit between neighboring geometrical elements as a harbinger of phase change.

It is hard to name another domain of physics in which so much power has flowed from such simplicity and propriety of means. The topic, like its tools, is a small diamond of the laboratory.

THE HUNGER FOR SALT: AN ANTHROPOLOGICAL, PHYSIOLOGICAL AND MEDICAL ANALYSIS, by Derek Denton. Springer-Verlag ($149.80). Going beyond a monograph, here is a treatise. Professor Denton is a distinguished experimental endocrinologist, director of the Florey Institute in Melbourne. There he has worked for 30 years with the docile sheep, surgically replumbed to offer answers to the interplay of need and supply, cue and behavior, in the intake and outflow of sodium ion. On this firm foundation of artful experiment and dauntless technique he has built a book with 10,000 citations, since he is as much at home in the library as he is in the surgery and the metabolism cages.

The topic is recognizably wide. The initial lines of text recall how Humphrey Davy "bounded about the room in ecstatic delight" once his great battery had isolated the first tiny globules of the two main alkaline metals. Sodium is the "main cation of body fluids," as salt flavors all the world's ocean. Inland of the salt-charged sea breeze, land life must somehow find sodium where only traces exist. How that happens, particularly with our own species, with the physiological background resting on the work with sheep and rats, and on clinical studies in the hospital and afield, fills 650 pages of a readable, reflective and marvelously catholic work.

Consider the relevance, and yet the unexpectedness, of documented and appraising accounts of cannibalism, of pica, an appetite for strange substances, of the cafeteria experiments, with sheep or children free to choose an offered variety of foodstuffs, of the salty diet of farming villages of Honshu, full of soy

WINE & FOOD

Cabernet Sauvignon

Cabernet Sauvignon is the unrivaled King of California wine grapes.

The distinction is well deserved for it was this magnificent varietal that first put California on the world map of wine. A native of Bordeaux, the wines produced of this grape have, for centuries, been among the world's finest and most enduring.

WINE

As with all heartier red wines, Cabernet Sauvignon must be properly aged to be at its best. When fully mature, a deep aromatic bouquet typically is followed by round, full flavors suggesting berries and herbs, warm earthy flavors, rich in varietal character. Velvety smooth in texture, it is a gentle giant of a wine.

Recognized for its excellence, our gold medal winning 1977 Proprietor's Reserve Cabernet Sauvignon

is classic in its composition and sufficiently aged to be entirely enjoyable now.

FOOD

Traditionally, a wine as full bodied and flavorful as Cabernet Sauvignon is served with dishes of beef or venison, hearty stews, cheeses and rich pastas.

In our home, we enjoy Cabernet with a wide variety of foods. One I'm especially fond of is our own *Eagle Ranch Roast*, a wine country dish developed by my wife Vicki. Generously seasoned with olive oil, garlic and rosemary, this complex layering of flavors would overpower a wine of lesser body and character.

If you would enjoy sampling some of Vicki's recipes and learning more about wine, please write to the address below for your free brochure. "Wine and Food: Cabernet Sauvignon" a collection of family favorites uniquely compatible with this King of California Wines.

Discover Sebastiani Vineyards, a family winery in the Sonoma Valley.

Sam J. Sebastiani

Sam J. Sebastiani

Sebastiani Vineyards, Dept. G, P.O. Box AA, Sonoma, CA 95476

sauce, pickles and miso, and of the role of fire in the entry of hominids into the Temperate Zone. It was not warmth that counted but glow, with its ability to break the rigid circadian rhythm that then required, it is held, a strict 12-hour duration of sleep throughout the tropical night.

Or look at the pictures. They include a wild kangaroo at twilight in the Snowy Mountains, visiting the filter-paper cafeteria blocks that offered a choice of water or four distinct salts. (Sodium was the runaway favorite.) Five thousand camels are shown being assembled in Algeria for the 600-mile march south across the central Sahara to fetch salt from the ancient mines of Mali. The modified sheep of the Melbourne laboratory are here, their saliva diverted, an adrenal or a kidney carried now in a fold of neck skin convenient for valving and sampling. Phosphate is rarer than sodium; carnivores, Lapland reindeer and the cattle of the King Ranch alike seek it by long chewing on the bones they find on the deficient pasture. A series of autoradiographs that record the sites of specific protein synthesis is presented to illustrate the future direction of endocrine studies.

The still uncertain implications for medicine are worth summary; that they are tentatively and prudently advanced must go without saying. The measured output of essential sodium, like the input, varies among peoples from the extreme salt-free life of the fierce Yanomamö of the Amazon to the salt-doting Japanese towns of the north by a neat factor of 100. Some effects appear at these extremes, although they are only marginally clear. Over a wide range of intake, however, the sodium content of the human body remains in balance, stabilized by diverse feedback loops. The hormonal control is clear, if in detail we understand it only in part.

Animals are led to take required salt by instinct, responsive to its taste and not to a learned sense of well-being. A conjecture suggests that the signal of balance is the salt content not of the body as a whole but of specialized neurons that somewhere determine the taste for salt. Above this physiological level the learned response that most complex societies have nurtured encourages salt consumption beyond need by up to two orders of magnitude. The system adjusts, but at a cost of hormone manufacture up by a large factor, and by other consequences we may not yet perceive. One suggestion is strong: those Japanese villagers have an endemic high blood pressure. They die by cerebrovascular accident at a much higher rate than other populations. The matter is not resolved; epidemiology still offers "the Scottish court verdict: not proven." Nevertheless, a preventive approach might be a very "good investment" in a situation where a fifth of the people in the developed countries face circulatory-system hazard.

THE SUN, OUR STAR, by Robert W. Noyes. Harvard University Press ($20). If ever we are to understand our past and our future, we must learn to look at the ancient sun in new ways. That is the lesson a reader takes first of all from this well-written account of solar physics and astronomy today, done in satisfying depth without one equation, although with plenty of graphs and diagrams.

The new ways enrich this tale, which by no means neglects the basics of spectral lines and atomic abundances, of eclipse tracks and sunspot cycles. It is something of a surprise to see a solar telescope on the icy plateau at the South Pole itself. The location is natural enough: there in midsummer the sun never sets. Through five 24-hour days an international team watched the sun uninterruptedly. (In Christmas week the weather at the pole is favorable: the temperature rises often to a summery 30 degrees below zero F.) The aim is to record in detail the velocity patterns of the solar gases averaged over a large area of the sun's surface. Single-wave trains can be followed there for days on end. The results amount to solar seismology, the study of oscillations that extend deep into the sun.

For a long time we have recorded the "solar constant"—the total radiation arriving at the earth from the sun—from the thin air of the best mountain stations. That air is clear but not clear enough; in 1980 the solar input was measured from orbit, beyond the whole of the atmosphere, with an unprecedented accuracy of one part in 10,000. The sun was much mottled, and the solar constant went down markedly once a large sunspot group rolled into view. Before that natural-enough result we could not know whether some slight general brightening might not compensate for the obvious dark spots.

Magnetic fields on the sun have been known almost from the beginning of the century. Sunspots, of course, are magnetic. Today we detect the looping and bridging fields on the solar surface at every scale: the high prominences, the pinpoint X-ray spots, the vast coronal holes out of which plasma currents stream into space, to generate magnetic storms and the subtle aurora once they engage the earth's own magnetic regime. The flares, sudden intensely radiating patches, are magnetic lightning, so to speak, on a solar scale. Although we do not yet understand the mechanism, the energy spent by flares must arise from a magnetic store established by the drifts and twists of the conducting solar plasma, as on the earth our Lilliputian flashes arise from an electrostatic store set up by insulating winds laden with charged droplets.

Inside the sun we cannot see. His birth and death are treated as stellar evolution foresees, supported more by the regularities in distant stars than by any special view of the sun. Here there is little news, and it is bad. The neutrino observatory, that tank watched so well by the radiochemists deep in the South Dakota lode, is our one channel into the solar depths, the only way we know to exploit the closeness of our star for information from its central nuclear furnaces. The result, however, does not fulfill the expectations of the theorists: the signal is small. Professor Noyes prudently sees the question as being unresolved; there is a growing feeling that the mystery is in fact an illusion. (Recent replication of the difficult nuclear experiments on which the input data for the calculations are based does not confirm the older results with the expected precision.) One inference from our privileged view of the sun has been carried out to the remote stars: a dozen years of careful monitoring of characteristic strong coronal emission lines has shown unmistakable starspot cycles, with the more mature stars displaying regular activity like the sun's.

A chapter is devoted to solar physics at the receiving end, that is, to the effects of solar energy on the earth. The suspicions that solar variations affect the earth's climate are well drawn, with the conclusion that while we do not understand how it can be, the evidence grows "that Nature is able to do what scientists cannot yet explain." It may be that this problem deserves priority beyond any other in astronomy and meteorology.

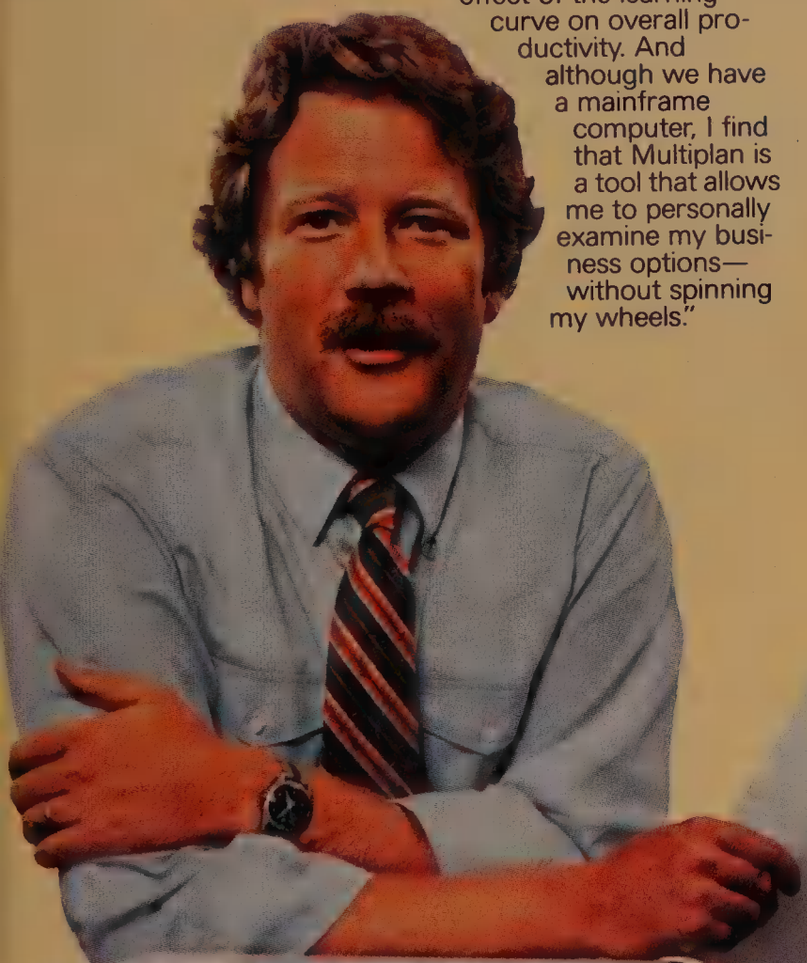
Text and illustrations join to set out a topic whose importance is as evident as its growth—and its incompleteness. Only one frustration mars a reader's pleasure in this straightforward and authoritative updating: scarcely any references are given (there is just one, in the preface), not even to the sources of the many figures and photographs, which much injures the welcome impression of timeliness and completeness. "Courtesy of..." phrases are simply not enough: we deserve to know who, how, when.

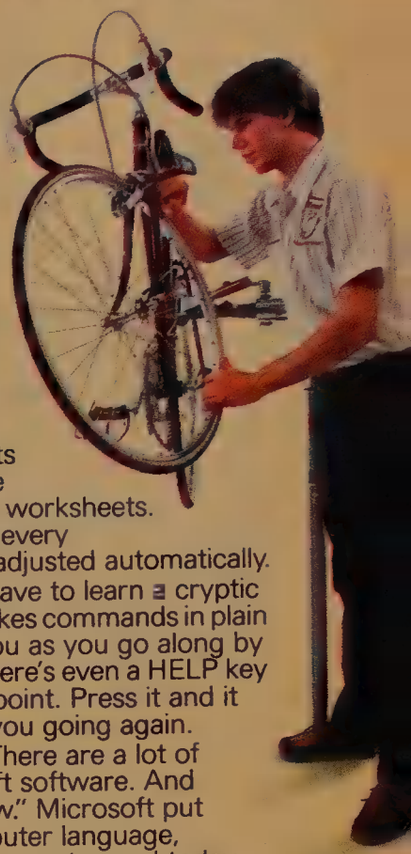
BLUEPRINTS: TWENTY-SIX EXTRAORDINARY STRUCTURES, by Christopher Gray, Patricia Brown and John Boswell. Simon and Schuster ($19.95). The Wright Flyer and the VW Beetle, the Hoover Dam and the Pentagon, the elephant-shaped house at Margate, N.J., and the Biltmore, a Loire chateau raised near Asheville, N.C., for a gilded Vanderbilt exemplify the extraordinary structures of this large paperback. Blueprints are reproductions of line drawings presenting the structure in simple projection. Characteristically they are works of the disciplined imagination:

"Microsoft® Multiplan™ allows me to explore more alternatives in less time. I think it leads to better management decisions."

Edward R. Schwinn, Jr.
President, Schwinn Bicycle Company

"Multiplan software helps me make better business decisions. It's as simple as that. With Multiplan on my microcomputer, I can explore a number of alternatives faster. That's really important to me in planning and tracking my totally hand-crafted Schwinn Paramount Bicycle. As with anything handmade, cost control and product planning are major considerations. For instance, Multiplan lets me study the effect of the learning curve on overall productivity. And although we have a mainframe computer, I find that Multiplan is a tool that allows me to personally examine my business options—without spinning my wheels."





A tool for decision makers. Microsoft Multiplan is a microcomputer software program that can help you too. In many ways. You can analyze cash flow. Plan budgets. Forecast income. Manage production. Multiplan allows you to set up an electronic worksheet for whatever your need may be. And, it lets you explore alternatives faster because when you make one change you immediately see the results of that change throughout the worksheet and on any related worksheets. Just change one number and every number that depends on it is adjusted automatically.

In plain English. You won't have to learn a cryptic language to use Multiplan. It takes commands in plain English. And, it will prompt you as you go along by telling you what to do next. There's even a HELP key in case you need help at any point. Press it and it gives you information to get you going again.

Backed by our knowhow. There are a lot of good reasons to use Microsoft software. And among the best is "knowhow." Microsoft put BASIC, the first microcomputer language, on the very first microcomputer, and today, Microsoft software is running on well over a million computers. Worldwide.

See it in action. If you'd like to explore more alternatives yourself, ask your local computer dealer for a demonstration of Microsoft Multiplan. Multiplan is available for most popular personal computers, including Apple,® the IBM® PC and computers that use CP/M-80® or MS™-DOS operating systems. It's a good management tool for building bicycles. It's also a good management tool for building businesses. Like yours.

BETTER TOOLS FOR MICROCOMPUTERS

MICROSOFT™

MICROSOFT CORPORATION, 10700 NORTHUP WAY, BELLEVUE, WASHINGTON 98004

Microsoft is a registered trademark, and Multiplan, MS, and the Microsoft logo are trademarks of Microsoft Corporation. Apple is a registered trademark of Apple Computer, Inc. IBM is a registered trademark of International Business Machines, Corp. CP/M-80 is a registered trademark of Digital Research, Inc.

they are made in advance, tools for the task in hand, "links between idea and reality." As engineering gets more complex, blueprints grow voluminously, but interest, even as documentation, inheres in very few. The details of fastenings and piping are useful mainly to those who maintain and repair. Only the most general of views become documents of value, and even they are often lost in the disposal of unwanted bulk.

This delightful compilation has searched out blueprints in such variety that some are sure to appeal to any reader caught by the technical. The large contrasty pages of white-lined images on a deep blue ground are graphically attractive, their meaning quite apart. Each example is here given less abstract representation by a couple of excellent large photographs to display the completed project, so that the book is something of a self-teaching introduction to technical drawing as well.

The first entry is winning. It is the Statue of Liberty; we see a view from the harbor, an aerial view of the island with its star fort, and two blueprint pages of Mlle Liberty. They are revealingly intimate, with sections at each landing, and above all an elevation that displays her iron skeleton, so unexpected and so necessary a gift from M. Eiffel himself. These blueprints are not original construction drawings; they were made in 1942 as insurance against bomb damage, since the old plans were then beyond reach in occupied France.

Quite a few other examples share this slight loss of authenticity, having been prepared after the fact. The contoured site plan of the Hoover Dam, the tower of the Empire State rising high and thin across four big pages and the elegance of a hyperbolic cooling tower are striking as well. The cooling tower is accompanied by views of Three Mile Island. It may take a generation for our public to look at those beautiful forms again without a sense of foreboding, "majestic structures... tarred with a brush they do not deserve."

Even the big Blueprint Company of New York no longer makes blueprints proper. Since the 1930's drawings have been reproduced mainly by the dry diazo process and others, black line in one step and no shrinkage. The word "blueprint" has become a metaphor for any detailed plan and will long outlive the simple and beautiful blue artifacts first made by Sir William Herschel in about 1840. The compilers of this visual enchantment must be confirmed New Yorkers; they have located the Pentagon with the accuracy of one who would remark that Coney Island is along the New Jersey shore.

GIRAFFES, BLACK DRAGONS, AND OTHER PIANOS: A TECHNOLOGICAL HISTORY FROM CRISTOFORI TO THE MODERN CONCERT GRAND, by Edwin M. Good. Stanford University Press ($29.50). The conveyor belt bears baby grands into the distance. We see the assembly line of the Yamaha plant at Hamamatsu, today the largest producer of pianos in the world, however unfamiliar to Chopin and Fats Waller. Japan now makes about 40 percent of the nearly one million pianos made each year around the world, about what the U.S. and the U.S.S.R. manage together. Integrated, automated without the neglect of craft skills where they are indispensable, the Yamaha firm (Nippon Gakki Co., Ltd.) modestly sees itself more as the Toyota than as the Rolls of the modern pianoforte. It ought to come as no surprise that the well-made iron frames of Yamaha pianos are cast by the vacuum-shield mold process, first developed to improve detail in the admirable motorcycle engines made by the same company. It is just a century since pianos came to Meiji Japan.

This lighthearted but careful book tells how over three centuries the modern piano came to be, more as an example of the development of the piano as a machine than as part of social history or the history of music. Design, materials, manufacture all play distinct roles. The sources are plentiful, including a large scholarly literature, an anecdotal richness, the records of companies and tradesmen, and above all the archaeology of the pianos themselves in collections worldwide. Pianos are not sealed sites, in the archaeologist's jargon, and the instrument preserved as Beethoven's own might have been repaired and modernized since the composer's time.

Bartolomeo Cristofori served a musical Medici prince at Florence as harpsichord maker, a man of recognized talent. Although the big harpsichord could hold its own in volume with the ensembles of the day, instrumental or voice, it lacked dynamics. Only by choosing an alternate keyboard could the player vary either loudness or timbre. The prince himself set the craftsman to improving the harpsichord. In 1700 Cristofori completed a "new invention that produces *piano* and *forte*." The instrument was a harpsichord save in one vital respect: the new action, the mechanism that actuated the strings, flung a hammer up at the string with a speed smoothly depending on the force of the key stroke. The device could be heard, unlike the little tinkly clavichord, and it offered the dynamics the big twanging harpsichord lacked, a new counterpart to voice and bowed strings.

In his time Mozart composed for the classical piano; in 1777 he praised at length the instruments made by Stein at Augsburg: "The tone is always even. [My last sonata] sounds exquisite on Stein's pianoforte." By the close of the 18th century the classical piano had decisively displaced the harpsichord. Wide pitch range and big volume were newly desirable. Beethoven and his times—concert tours, a new public, a new forcefulness in music—were reflected in, although not directly responsible for, the new pianofortes. The finest London maker, Broadwood, sent the deaf Beethoven a new grand with a big tone. It came too late to be heard.

Pitch range and volume grew. Long and heavy strings meant much stored energy. Strong players such as Beethoven and Liszt broke strings in bunches. The tension began to warp the wood frames, braced as they might be. Between 1820 and 1840 or so iron entered and dominated the piano. Cast-iron frames held the tens of tons of total stress required by more than eight octaves of loud strings. Steel wire, annealed and drawn through dies of diamond and ruby, became available sometime before 1840; piano wire remains to this day an engineering standard of steel uniformity and strength.

The modern grand piano appeared in about 1860 under the name of Steinway & Sons, American innovators and world leaders for generations. The upright pianos are more diverse, if commoner, and they settled down somewhat later. Grotesque vertical forms like the giraffe were long gone. (In 1904 the American manufacturers collected and made a pyre of all the big black square pianos they could find.) Piano evolution had reached a plateau.

New urea-based glues, Teflon bushings at many bearing points and plastic keys mark the main novelties over the past century. The player piano won a brief ascendancy in the 1920's, to be all but expunged by the phonograph and the radio. Electronic pianos, both those that seek to resemble pianos in tone and those that do not, have budded but not yet flowered. The concert grand and the home piano differ now more than ever; so does their music.

The volume surveys its unexpected landscape without pretense and with careful attention to evidence. The place of acoustical science seems very small; the great Helmholtz had as students both Theodore Steinway and the inventor of an ingenious but never-popular new keyboard. Little seems to have come to this story, however, out of the laboratory, apart from materials. Indeed, a bit more physics would help the tale along. The notes and apparatus of the book are in general good, although one important page has been reduced to the ridiculous by an editorial oversight. An account of the piano action, not at all a simple mechanism, refers to part after part by the letters, from key *A* to jack spring *P*, but not one letter appears on the facing diagram. The editor responsible should be set to sanding a bushel of capstans, knuckles and wippens, whatever they are.

Take the 6-volume Microcomputing Library for only $2.95

when you join the Small Computer Book Club.

You simply agree to buy 3 more books—at handsome discounts—within the next 12 months.

This set of six carefully selected books—a $76.25 value—will be your basic reference source for small computing.

BASIC AND THE PERSONAL COMPUTER.
Thomas A. Dwyer and Margot Critchfield. Clear, down-to-earth introduction to programming in BASIC. Detailed examples and numerous illustrations explain the wide range of applications possible with personal computers. Also, many step-by-step examples of word processing, computer games, and computer simulations. **Publisher's Price $14.95**

USING CP/M: A Self-Teaching Guide.
J.N. Fernandez and R. Ashley. A complete introduction to the CP/M ("Control Program") software package used on many advanced microcomputers. How to use CP/M to copy files, edit, test programs, translate programs, and more—regardless of the programming language the microcomputer uses. **Publisher's Price $14.95**

MICROCOMPUTER GRAPHICS.
Roy E. Myers. Provides the essential mathematics and programming techniques you will need for computer graphics applications in BASIC, and shows you how to handle animated figures, peripheral equipment, colors and resolution, and many other topics. **Publisher's Price $11.95**

INTERFACING MICROCOMPUTERS TO THE REAL WORLD.
Murray Sargent III and Richard L. Shoemaker. The complete guide for computerizing your home, office, or laboratory. Tables, charts, and displays show you how to handle microcomputer software and operating systems; and how to monitor everything from the house lights to motors, switches, and display panels. Examples feature the Z-80 microprocessor. **Publisher's Price $14.50**

THE WORD PROCESSING BOOK: A Short Course in Computer Literacy.
Peter A. McWilliams. You'll see how to save up to $15,000 by buying a small computer with word processing capacities. Includes a brand name buying guide. **Publisher's Price $9.95**

EXECUTIVE COMPUTING.
John Nevelson. 25 business methods illustrated by computer programs, in a step-by-step casebook. You'll find details on pricing techniques...long-range planning...corporate data bases...and much more. **Publisher's Price $9.95**

SMALL COMPUTER BOOK CLUB is a division of the Library of Computer and Information Sciences, the oldest, largest and most respected computer book club in the world. The SMALL COMPUTER BOOK CLUB will keep you up-to-date with the latest developments in software, improvements in hardware, programming languages, and much more—all at handsome discounts.

So start enjoying the club's benefits today.

4 Good Reasons to Join

1. The Finest Books. Of the hundreds of books submitted to us each year, only the very finest are selected and offered. Moreover, our books are always of equal quality to publishers' editions, *never* economy editions.

2. Big Savings. In addition to getting the Microcomputing Library for $2.95 when you join, you keep saving substantially—up to 30% and occasionally even more. (For example, your total savings as a trial member—including this introductory offer—can easily be over 50%. That's like getting every other book free!)

3. Bonus Books. Also, you will immediately become eligible to participate in our Bonus Book Plan, with savings up to 70% off the publishers' prices.

4. Convenient Service. At 3-4 week intervals (16 times per year) you will receive the Book Club News, describing the Main Selection and Alternate Selections, together with a dated reply card. If you want the Main Selection, do nothing and it will be sent to you automatically. If you prefer another selection, or no book at all, simply indicate your choice on the card, and return it by the date specified. You will have at least 10 days to decide. If, because of late mail delivery of the News, you should receive a book you do not want, we guarantee return postage.

If the reply card has been removed,
please write to the **Small Computer Book Club**
Dept. Y-AB5, Riverside, N.J. 08075
to obtain membership information and an application.

The Riviera collection.

If you're an admirer of the fine art of car building and design, here are three exceptional works by Buick that you'll want to consider.

The luxurious Riviera. The performance-tuned Riviera T-TYPE. And the new and very limited edition Riviera Convertible.

Can't quite see buying Riviera purely out of appreciation for the art? Then consider the incredible list of equipment you'll be acquiring. Even in standard form, Riviera ranks as one of the world's most completely outfitted cars.

The Rivieras. If you can't decide which one would be best to start your collection with, buy all three. After all, why break up the set?

Official Car of the XXIIIrd Olympiad
Los Angeles 1984

BUICK

Wouldn't you really rather have a Buick?

Established 1845 SCIENTIFIC AMERICAN April 1983 Volume 248 Number 4

A Nuclear-Weapon-Free Zone in Europe

A proposal to ban nuclear weapons of all kinds from a strip of land on both sides of the East-West boundary in central Europe is viewed as a politically feasible approach to reducing the risk of nuclear war

by Barry M. Blechman and Mark R. Moore

The nuclear-arms-control talks now under way in Geneva between the U.S. and the U.S.S.R. are directed toward two goals: reducing the number of strategic, or long-range, nuclear weapons deployed by the two superpowers and reducing the number of intermediate-range nuclear weapons capable of reaching targets in Europe. Excluded from the negotiations is another major category of nuclear arms, namely tactical, or short-range, nuclear weapons. Of the roughly 6,000 nuclear weapons currently deployed in Europe by the military forces of the North Atlantic Treaty Organization (NATO) more than half are classified as tactical. The opposing forces of the Warsaw Pact are thought to have a comparable arsenal of nuclear weapons based on their side of the East-West boundary in central Europe. The short-range nuclear weapons deployed by both sides in Europe are intended to provide direct support to ground forces engaged in combat; hence they are also known as battlefield nuclear weapons. Included under this heading are nuclear warheads for short-range ballistic missiles (defined here as those with a maximum range of 150 kilometers), artillery shells (with a maximum range of 30 kilometers), antiaircraft missiles and explosive mines.

The large numbers of battlefield nuclear weapons based in Europe present a special problem in arms control. Be-

Editor's Note

The Independent Commission on Disarmament and Security Issues, which included the call for a battlefield-nuclear-weapon-free zone in its final report, published last June, was made up of 18 members, all of whom served in a private capacity. The chairman of the commission was Olof Palme, then a member of the Swedish parliament, chairman of the Swedish Social Democratic Party and former prime minister of Sweden. (Palme has since been reelected Sweden's prime minister.) The other commissioners were Giorgi Arbatov, full member of the Central Committee of the Communist Party of the U.S.S.R., deputy of the Supreme Soviet and director of the Institute of the U.S.A. and Canada, a division of the Academy of Sciences; Egon Bahr, member of the West German parliament, chairman of that body's subcommittee on disarmament and arms control and former minister for economic cooperation; Gro Harlem Brundtland, member of the Norwegian parliament, chairman of the Norwegian Labor Party and former prime minister of Norway; Jozef Cyrankiewicz, former prime minister of Poland and former president of the Council of State; Jean-Marie Daillet, member of the French parliament, vice-chairman of that body's defense committee and chairman of the Giscardist UDF party's defense committee; Robert A. D. Ford, special adviser on East-West relations to the Canadian government, former Canadian ambassador to Colombia, Yugoslavia, Egypt and the U.S.S.R.; Alfonso Garcia-Robles, ambassador and chairman of the Mexican delegation to the United Nations Committee on Disarmament and former foreign minister of Mexico (co-winner of the 1982 Nobel peace prize for his role in the creation of a Latin-American nuclear-weapon-free zone); Haruki Mori, former Japanese ambassador to Britain and to the Organization for Economic Cooperation and Development and former vice-minister in the Japanese Ministry of Foreign Affairs; C. B. Muthamma, Indian ambassador to the Netherlands and former ambassador to Ghana and Hungary; Olusegun Obasanjo, general in the Nigerian armed forces, member of the Council of State, Distinguished Fellow of the University of Ibadan and former head of state of Nigeria; David Owen, member of the British parliament and former secretary of state for foreign and Commonwealth affairs; Shridath Ramphal, secretary general of the Commonwealth and former foreign minister of Guyana; Salim Salim, minister of foreign affairs of Tanzania; Soedjatmoko, rector of the United Nations University in Tokyo and former Indonesian ambassador to the U.S.; Joop den Uyl, member of the Dutch parliament, deputy prime minister of the Netherlands, leader of the Dutch Labor Party and former prime minister; Cyrus Vance, former secretary of state of the U.S.

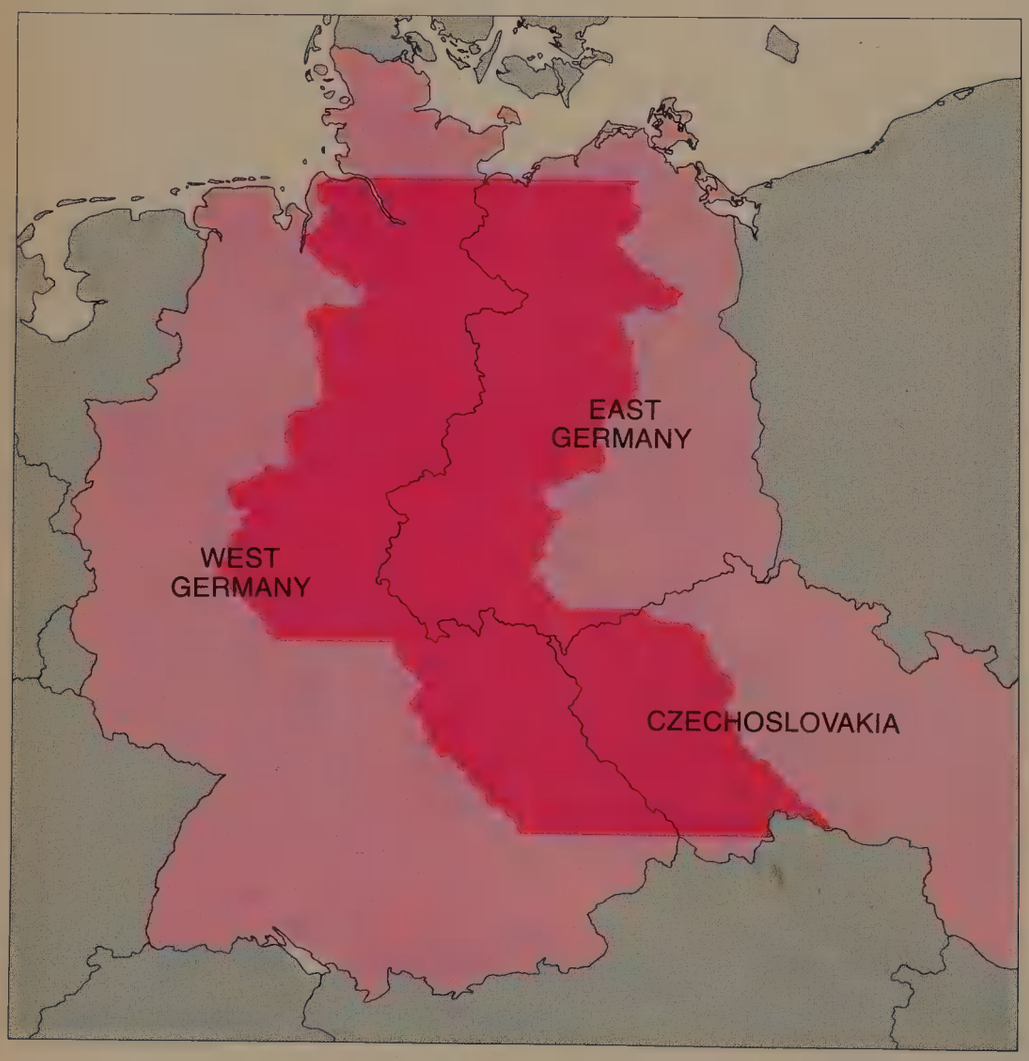
cause many of them are deployed in forward positions in central Europe, close to the front lines in any likely war, they stand a good chance of being overrun by invading forces. Given a choice between using the weapons or losing them, the defending side would presumably be under great pressure to fire them first, thereby escalating the conflict past the "nuclear threshold."

For that matter, in the kind of intense political crisis that would be likely to precede the outbreak of hostilities there could be considerable pressure on political leaders to delegate in advance to military commanders in Europe the authority to release battlefield nuclear weapons for use under certain circumstances. Whether any American president or European leader would in fact make such a decision is questionable; nevertheless, the pressure to do so would be real and the possibility cannot be dismissed. Even leaving aside the possibility of such a deliberate decision, the problem of maintaining effective control of battlefield nuclear weapons would be difficult in wartime. Given the confusion and misinformation that inevitably accompanies large-scale military operations, the specific location of nuclear-weapon sites could be a crucial factor in determining whether or not a decision is made to initiate nuclear war.

A novel plan for dealing with this problem was put forward last year by the Independent Commission on Disarmament and Security Issues, an international group of present and former government officials formed in 1980 under the chairmanship of Olof Palme, who was then the former prime minister of Sweden. (He is now again prime minister.) The commission, the members of which were from Eastern, Western and Third World nations, included Cyrus Vance, former secretary of state of the U.S., and Giorgi Arbatov, a member of the Central Committee of the Communist Party of the U.S.S.R. In a report published in June of last year the commission called for the establishment of a "battlefield-nuclear-weapon-free zone" in Europe, together with other "measures to strengthen the nuclear threshold and reduce pressures for the early use of nuclear weapons."

According to the commission's report, *Common Security: A Blueprint for Survival,* the proposed nuclear-weapon-free zone would start with central Europe, perhaps as a strip 150 kilometers wide on each side of the border of West Germany with East Germany and Czechoslovakia. In time the zone would be extended, reaching ultimately from the northern to the southern flanks of the two alliances. Nuclear weapons of any kind would be banned in the zone.

The concept of a nuclear-weapon-free zone in Europe is not new. Nevertheless, the commission's proposal differs from all the previous proposals in several important respects, not the least of which is that it is based on the characteristics of a particular category of nuclear weapons rather than on some other criterion, such as the territorial extent of the nations involved. In our view the commission's plan is a practical and politically feasible way of immediately reducing the risk of nuclear war in Europe. In this article we shall summarize the arguments for a battlefield-nuclear-weapon-free zone in central Europe. In particular we shall explain why we believe this particular plan could succeed where all the others have failed.



PROPOSED DENUCLEARIZED ZONE in central Europe would be rid of all nuclear weapons, thereby reducing the risk that in the event of a conventional military attack the defending side could be pressured into firing its battlefield nuclear weapons first in order to keep them from being overrun by the invading forces. According to the report of the Palme commission, the exact geographic definition of the zone "should be determined through negotiations" between the interested parties. For illustrative purposes, however, the commission suggested a zone consisting initially of a strip of land 150 kilometers wide on each side of the border of West Germany with East Germany and Czechoslovakia (*dark color*). Unlike earlier proposals of this type, nuclear weapons could still be stationed in parts of all three countries (*light color*).

Most nations have already agreed that certain places should remain free of nuclear weapons: Antarctica, outer space and celestial bodies, the seabed and Latin America are all subject to treaties that forbid the deployment of nuclear weapons within their domain. Proposals have also been made to ban nuclear weapons in the Middle East, South Asia, the South Pacific and Africa, and in Europe as well. Each of these proposals has been the subject of at least some discussion in various international forums. Nevertheless, the history of efforts to create nuclear-weapon-free zones has yielded far more proposals than results. Indeed, the proposals often appear to be diplomatic maneuvers for political advantage rather than genuine attempts to reduce the risk of nuclear war.

The first and most detailed plan for a nuclear-weapon-free zone in Europe was put forward in 1957 by Adam Rapacki, then foreign minister of Poland. The Rapacki plan, as it came to be known, proposed that the governments of Poland, Czechoslovakia, East Germany and West Germany agree not to manufacture, maintain or allow the installation of nuclear weapons within their borders. The denuclearized status of the zone would be guaranteed by the existing nuclear-weapons states (then the U.S., the U.S.S.R. and Britain), and a

system of ground and air surveillance would be set up to ensure compliance. As the Rapacki plan was originally proposed, it did not envision a formal treaty establishing the zone; it relied instead on unilateral declarations by the interested governments.

The political objectives of the Rapacki plan seemed obvious to the Western nations. The plan was proposed at a time when the governments of the NATO countries were actively discussing the possibility of offsetting the superior conventional (non-nuclear) military forces of the Warsaw Pact by means of a "nuclear shield." Indeed, the decision to take this approach was made soon afterward, at a North Atlantic Council meeting in December, 1957. As a result the NATO countries viewed the Rapacki plan as an attempt to weaken their military position in central Europe, and they rejected it.

The Polish plan was twice revised and resubmitted to the Western nations, in November, 1958, and March, 1962. The changes did not meet NATO's basic objections, and so they made little headway. Subsequent Eastern proposals for nuclear-weapon-free zones in Europe have concentrated on subregions of the continent. The Balkans and the Adriatic area have long been the subject of proposals, and so has the entire Mediterranean basin. The Nordic region, which Premier Khrushchev of the U.S.S.R. first suggested (in 1959) as a suitable zone to be denuclearized, has in recent years again received much attention from Russian officials.

Not long after Eastern proposals for the creation of nuclear-weapon-free zones were first being made, the U.S. developed a list of four criteria that such an initiative would have to meet in order to gain American support: (1) The proposal should be initiated voluntarily by states in the region; (2) the proposal should include all the states of the region, or at least all those with significant military power; (3) the existing military balance and security arrangements should not be disturbed; (4) the agreement should include provisions for the verification of compliance.

In the case of Latin America the U.S. evidently found that the proposal to create a nuclear-weapon-free zone fitted, or could be modified to fit, these criteria. On the other hand, the U.S. and other Western governments have not found past proposals for the creation of nuclear-weapon-free zones in Europe to be suitable subjects for negotiation. Some of the objections have been technical: given the long-standing opposition of the U.S.S.R. to on-site inspection, the task of ensuring the withdrawal and continued absence of nuclear weapons from a region already densely populated with them has seemed difficult if not impossible.

	WEAPON	FIRST DEPLOYED	RANGE (KILOMETERS)	MAXIMUM YIELD (KILOTONS)	NUMBER OF LAUNCHERS
NATO	LANCE	1973	115	50	90
	HONEST JOHN	1953	38	20	90
	M109 155-MILLIMETER GUN	1964	18	2	1,455
	M110 203-MILLIMETER GUN	1962	16	1	390
	PLUTON (FRANCE)	1974	120	10	42
WARSAW PACT	FROG/SS-21	1950's/1977	75/155	200	375
	SCUD-A	1957	85	1	250
	203-MILLIMETER GUN	1979	31	?	150
	240-MILLIMETER GUN	1979	30+	?	150

BATTLEFIELD NUCLEAR WEAPONS currently deployed on both sides of the East-West boundary in central Europe include the short-range ballistic missiles and artillery pieces listed in this table. Included for the forces of the North Atlantic Treaty Organization (NATO) are all such systems assigned to the European theater of operations. (France, although it is a member of NATO, is not an integrated part of the military alliance.) Included for the forces of the Warsaw Pact are all such systems stationed outside the U.S.S.R. and in the western military districts of the U.S.S.R. All the missiles shown are capable of delivering only one nuclear warhead each. Not shown are nuclear explosive mines and nuclear-armed surface-to-air missiles, both of which are sometimes classified as tactical nuclear weapons. Also excluded are the thousands of longer-range nuclear weapons deployed on land or at sea within range of European targets.

The fundamental Western objections to nuclear-weapon-free zones in Europe, however, have arisen from two other factors: the basic asymmetry in the geopolitical characteristics of the U.S. and the U.S.S.R., and the key role nuclear weapons have come to play in Western military strategy and in political relations among the NATO partners. Given the weight of these considerations, Western officials have tended to see Eastern proposals for the creation of nuclear-weapon-free zones in various parts of Europe as being largely propagandistic, meant to encourage dissension within the Western European nations and to disrupt political bonds among the members of the alliance. NATO's reservations must be fully understood if realistic measures to reduce the risk of nuclear war in Europe are to be designed.

Throughout the history of the Western alliance officials and ordinary citizens in the Western nations have debated the degree to which it would be effective and prudent to rely on nuclear weapons to maintain the peace in Europe. Advocates of conventional defenses argue that the surest route to stability in Europe is to maintain conventional military forces strong enough to persuade the leaders of the U.S.S.R. that in the event of conflict they would be defeated on the battlefield and thus face a choice between escalating to nuclear war or accepting defeat. If the Russian leaders believed such a choice would be the likely result of conflict in Europe, it is argued, they would decide against war, choosing instead to favor stable relations between the Eastern and Western nations. Support for this position has led to recurrent efforts to strengthen NATO's conventional forces.

For many other people in the Western nations, however, the threat of nuclear war, and the forces necessary to make the threat credible, have always been the essence of NATO's strategy. From this point of view what persuades the leaders of the U.S.S.R. to seek stability and avoid war in Europe is not a calculation that they would be likely to lose such a conflict but rather a general perception that if events were to get out of hand, the results of any military engagement could be disastrous. Indeed, people holding this view argue that NATO's military forces should be designed to avoid a situation in which the results of conflict become calculable. If the outcome can be predicted, they maintain, it may become acceptable under certain conditions. It would be far better, they argue, to enhance deterrence by inculcating the overriding perception that the inexorable result of conflict would be catastrophe.

Those holding this view seek to make clear to the leaders of the U.S.S.R. that any overt aggressive act in Europe could easily deteriorate into conflict and eventually escalate into an all-out nuclear war between the U.S. and the U.S.S.R. In short, they seek to maintain a position in which the ultimate danger of confrontation in Europe—nuclear holocaust—is made as stark as possible. No improvement in conventional forces can serve this purpose. Indeed, from this point of view the stronger conventional defenses seem, the greater the risk is that the leaders of the U.S.S.R. might err in their calculations and commit aggression, in the

belief that such defenses had reduced the risk of nuclear war.

For this reason, among others, NATO has never fulfilled its stated objective of strengthening its conventional forces. It is also the reason NATO declares explicitly that, if necessary, it would initiate and escalate a nuclear war in Europe. Furthermore, it is the reason NATO maintains thousands of tactical nuclear weapons in Western Europe, is now deploying a new generation of intermediate-range nuclear weapons and regularly expresses concern about the balance between the strategic nuclear forces of the U.S. and of the U.S.S.R.

There is yet another reason, however, for this emphasis on nuclear defense. For almost 40 years American strategists and political leaders have believed the most dangerous threat to American interests is represented by the U.S.S.R. The U.S.S.R. is perceived not only as threatening U.S. security in military terms but also as challenging in subtler ways the values of American society. As a result, when American strategy is viewed in geopolitical terms, its essence throughout this period has been to contain both the military power and the political influence of the U.S.S.R. within the Eurasian land mass. Most important, U.S. strategy has sought to limit the influence of the U.S.S.R. in the parts of Europe beyond the belt of nations occupied by the Russian armies in 1945. The chosen vehicle of the strategy, of course, is NATO.

In the struggle for influence in Western Europe between the two superpowers the U.S. is at a distinct geographic disadvantage. The military power of the U.S.S.R. is an immutable presence felt strongly in the capitals of Western Europe. In contrast, regardless of the vast economic and political stake of the U.S. in the continued independence and pro-American orientation of Western Europe, the U.S. security commitment and military presence in Europe is now, and always will be, perceived as somewhat transient and unreliable. Western Europeans recognize that the U.S. always has the option of withdrawing beyond the oceans that have sheltered it in the past. Such Western European fears can never be fully allayed. Indeed, they often are aggravated when the U.S. shows evidence of having one of its recurrent bouts of neoisolationism.

To offset the asymmetry in both the perception and the reality of the American and Russian military presence in Western Europe, U.S. decision makers have made continual efforts to strengthen the links between U.S. forces stationed in Western Europe and overall U.S. military power, and to reassure allies (and caution adversaries) about the durability of U.S. commitments. These efforts take various forms, but the crucial components have been steps to "couple" U.S. strategic nuclear forces with ground forces deployed in Europe.

Western strategy in Europe is based on the concept of escalated responses to aggression, beginning with conventional counterattacks if the Warsaw Pact attack is limited to conventional forces but including the possibility of the initiation and continued escalation of nuclear war. This declared position is given weight by the deployment of short-range nuclear weapons in Europe not far from the putative front lines; these weapons, which are designed to be used against military forces on the battlefield, would in many scenarios be the first to be used if a decision were made to cross the nuclear threshold. Because they are of comparatively small yield and would be employed against military targets, they are believed to make the threat of initiating nuclear war more credible.

To add further credibility to NATO's nuclear threat, the alliance plans to deploy U.S. intermediate-range missiles in Western Europe that are capable of striking targets in Russian territory. If the use of battlefield nuclear weapons did not lead to a termination of the conflict, then the intermediate-range weapons could be launched, still against military targets. This step would presumably be less difficult than the next one: the engagement of the central strategic forces of the U.S. Even so, nuclear strikes on Russian territory by U.S.-controlled missiles would almost certainly lead to a Russian response against the U.S., thereby precipitating the involvement of the two sides' central strategic forces. The idea is that since the leaders of the U.S.S.R. would recognize this sequence of escalation, they would be deterred from taking the first step on the escalation ladder. In theory deterrence is strengthened to the extent that the eventual involvement of U.S. strategic nuclear forces is made credible.

In one sense this sharing of risk is the ultimate psychological and practical bond of the NATO alliance, at least with respect to its security policies. By promising to resort, if necessary, to strategic nuclear forces in defense of Western Europe, thereby accepting the likelihood of retaliatory Russian attacks on U.S. territory, the U.S. has injected itself symbolically as a full partner in the fate of Europe, thereby offsetting the geographic isolation of the U.S.

In the light of this reasoning, the flaws in Eastern proposals to denuclearize parts of Europe, as seen in NATO capitals, should be apparent. As traditionally proposed, a nuclear-weapon-free zone would deprive Western Europe of a vital link between its conventional defenses and the U.S. strategic nuclear deterrent, thus emphasizing existing asymmetries both in the two sides' geographic positions and in the balance of conventional military forces. This reasoning applies primarily to proposals (such as the Rapacki plan) for the creation of nuclear-weapon-free zones in central Europe, the subregion at the heart of all considerations of the defense of Western Europe. Similar proposals for other subregions share some of the

U.S. LANCE MISSILE is seen being readied for a test launch from its self-propelled erector-launcher vehicle. The liquid-fuel rocket can carry either a nuclear warhead or a conventional one. Its maximum range is about 115 kilometers. The Lance has replaced the Honest John as the standard U.S. battlefield nuclear missile. Other NATO nations have both the Lance and the Honest John in their arsenals. All nuclear warheads for the missiles are in U.S. custody.

U.S. M110 203-MILLIMETER GUN is shown being test-fired with a non-nuclear shell. The eight-inch howitzer is also capable of firing a nuclear shell. Its maximum range is approximately 16 kilometers.

The gun can be mounted either on a self-propelled vehicle, as is seen here, or on a towed chassis. The structure at the rear absorbs the recoil. The U.S. has supplied M110's to most of its NATO allies.

same political and military deficiencies, however, and in any case they are regarded as unhelpful precedents that might make a central European nuclear-weapon-free zone harder to reject.

From NATO's standpoint, therefore, the weaknesses of nuclear-weapon-free zones, as they have been traditionally proposed, are as follows. First, the initiatives typically are not coupled with proposals for restraints on conventional forces. This means that the ability of NATO to offset the Warsaw Pact's greater conventional forces with the threat of nuclear escalation would be undermined. Of course, NATO could mobilize its greater economic resources to compete more effectively in terms of conventional military forces, but that would be expensive and seems unlikely in the face of the traditional reluctance of the Western democracies to sustain heavy military burdens in peacetime.

Moreover, even if more effective conventional defenses were to be mounted, there would remain the question of the extent to which the lessening of NATO's nuclear threat would reduce the ability of the Western European nations to resist Russian political pressures and even military aggression. Although it would be impossible to prove, the fact that international relations in Europe have evolved into a fairly stable configuration over the past 35 years suggests that NATO strategy may well have fulfilled its designers' military and political expectations. At a minimum it indicates the existing doctrine should not be modified without substantial reason to expect that alternatives would be equally successful and better suited to current circumstances.

Second, because the removal of nuclear forces from central Europe would suggest the decoupling of the American strategic nuclear deterrent from the defenses of Western Europe, it might be taken to presage the political decoupling of Western Europe from the U.S. Such a development, it is feared, would lead inevitably to greater Russian influence in, and eventual dominance over, European affairs. The fact that Russian territory was not included in proposals to denuclearize central Europe highlighted the problem. It meant in effect that if NATO accepted Rapacki's proposal or subsequent ones, it would be acquiescing in a diminution of the American presence in Western Europe without any commensurate diminution of the overpowering Russian presence. The political implications of such a poor bargain are striking. It would be widely interpreted as suggesting that decisive Russian influence in the affairs of Western Europe had been conceded.

One more concern enters NATO evaluations. In addition to its purpose of containing Russian power, the formal partnership of West Germany with other Western European and North American states in a military alliance was intended to help heal the wounds of World War II. Along with the various multinational economic institutions created in the 1950's, NATO fostered the reconciliation of Germany and its former enemies, making possible the development of stable relations in Europe and the avoidance of the national tensions that had led to disastrous wars three times in the span of a century.

The Rapacki plan would have placed East and West Germany (along with Poland and Czechoslovakia) in a special category. West Germany would no longer have been a full member of the alliance. Moreover, such a decision would not have been made by the Germans themselves but rather would have been imposed on Germany by outside forces. This would have emphasized Germany's special position and, it was feared, would have strengthened prospects for the emergence of nationalist forces within Germany that could jeopardize its political ties with other Western nations and eventually those between the Eastern and Western nations. In short, by singling out Germany, the Rapacki plan and other proposals for nuclear-weapon-free zones in central Europe raised fears that old tensions might be rekindled. Far better, in the eyes of many Western leaders, would be the continued integration of Germany with the rest of Western Europe and the gradual improvement of relations between the two alliances as a whole.

Thus the Rapacki plan and subsequent proposals for nuclear-weapon-free zones in central Europe were marred primarily by their political implications. Because they suggested a clear differentiation of risk between the U.S. and Europe, they were interpreted as being means for the erosion of American influence in Europe and the aggrandizement of Russian power. Because they singled out Germany for special

treatment, they were perceived as being likely to undermine the efforts that had been made to integrate West Germany into the Western community of nations. The current proposal of the International Commission on Disarmament and Security Issues for a battlefield-nuclear-weapon-free zone has been carefully designed to avoid these disadvantages.

The commission took careful note of the complex relations among the different types of military forces in Europe and urged a comprehensive approach to arms control as the only way to ensure the achievement of lower levels of forces and a reduced risk of war. The commission stated in its report that "a negotiated agreement for approximate parity in conventional forces between the two alliances... would facilitate reductions in nuclear weapons and a reordering of the priority now accorded to nuclear arms in military contingency planning."

Accordingly the commission began its recommendations for Europe by urging the convening of a conference of the foreign ministers of the NATO and Warsaw Pact nations that have been negotiating in Vienna for the past 10 years for an agreement on conventional forces. It called on the foreign ministers to complete a first-phase agreement that would guarantee parity between the two sides and make possible continued negotiations for substantial reductions in conventional forces. In the view of the members of the commission, the differences remaining between the two sides could be resolved fairly quickly if there were the political will to resolve them. The commission concluded that agreement to hold a meeting at the foreign-minister level could provide both the impetus and the opportunity to move things forward.

The commission also paid considerable attention to chemical weapons, urging the removal of such weapons from all Europe and proposing associated measures to assure both sides that such a prohibition was being observed. The latter measures would include on-site inspections on a "challenge" basis, meaning that each side would be allowed to inspect a certain number of facilities of its choosing on the other's territory each year, if it wanted to exercise the option. This is an important and creative proposal. That the Russian representative on the commission agreed to the inspection provision suggests a greater Russian responsiveness to Western demands for serious attention to the problems of verifying arms-control agreements than has been evident in the past. Russian responsiveness to concerns about verification was also evident in the discussion of the proposed nuclear-weapon-free zone.

The commission considered several alternative approaches to the nuclear component of European security problems, including the creation of nuclear-weapon-free zones defined by national boundaries. In the end, however, the members of the commission decided to pursue a functional approach, concentrating on specific categories of weapons that raise particular problems.

The commission called special attention to "the dangers posed by those nuclear weapons whose delivery systems are deployed in considerable numbers to forward positions in Europe." It urged the mutual removal of battlefield nuclear weapons from regions close to the East-West border through the negotiation of a battlefield-nuclear-weapon-free zone. To minimize verification difficulties the commission proposed that no nuclear munitions be allowed in the zone. It stated further that storage sites for nuclear munitions, maneuvers simulating nuclear operations and preparations for the emplacement of nuclear mines should also be prohibited. Finally, it suggested special rules should be negotiated covering the deployment in the zone of "dual purpose" artillery and short-range missiles that could be used to deliver nuclear munitions.

The commission believed it was best not to recommend specific boundaries for the zone, leaving that task to the parties concerned. For illustrative purposes, however, the report cited the possibility of a 300-kilometer zone centered on the East-West boundary [*see illustration on page 38*]. Such a zone would initially include parts of West Germany, East Germany and Czechoslovakia. Unlike the Rapacki plan and subsequent Eastern proposals, however, parts of all three nations would still be available for the deployment of nuclear weapons; this was a crucial distinction. As the zone was extended to the north and the south, depending on how its boundaries were set, parts of Austria, Hungary, Yugo-

RUSSIAN SHORT-RANGE MISSILES are transported across a muddy field on tracked erector-launcher vehicles. The missiles, designated SCUD-A by Western intelligence sources, can fire either a nuclear warhead or a conventional one. Their maximum range is

slavia, Italy, Denmark, Sweden, Norway, Finland and the U.S.S.R. could be added to it.

The commission also left it to the negotiators to define the specific provisions needed to ensure the parties' ability to verify compliance with the agreement. The members did agree, however, that such provisions "would have to include a limited number of on-site inspections in the zone on a challenge basis." Again the fact that the Russian representative agreed to include this phrase is significant. Although Arbatov, like all the other members of the commission, served in a private capacity, it is evident that his assent to the text—particularly the recommendations—was not given without higher authorization. The acceptance of on-site inspections in this context is consonant with the trend toward greater Russian receptivity to more intrusive means of verification.

Finally, the commission recommended some additional measures to further strengthen the nuclear threshold and reduce pressures for the early use of nuclear weapons. Most pointed was the recommendation of a majority of the commission that nuclear weapons that might appear more usable, particularly small nuclear weapons such as enhanced radiation weapons ("neutron bombs"), not be deployed in Europe. Although a specific recommendation was not made, a section of the report also discussed the special problems associated with nuclear-armed antiaircraft missiles and explosive mines, suggesting that the removal of these weapons from Europe could further strengthen the barriers against the initiation of nuclear war.

about 85 kilometers. Nuclear warheads for Warsaw Pact weapons are in Russian custody.

In short, the battlefield-nuclear-weapon-free zone proposed by the commission appears to escape the flaws associated with previously proposed nuclear-weapon-free zones. What would the implementation of the battlefield-nuclear-weapon-free zone accomplish?

For one thing, it would strengthen the barriers against the inadvertent or accidental initiation of nuclear war. It seems clear that nuclear war is not likely to result from a cold-blooded calculation of advantage during normal times; the forces of both sides are too large for either side to see any meaningful advantage to be gained. It is likelier that nuclear war would emerge from desperate decisions in an intense crisis, probably one in which the conventional forces of the two sides had already been engaged. Under such circumstances one side or the other might initiate a nuclear attack because it saw no other option for averting a catastrophic defeat, or because it had concluded that its adversary was about to escalate to the nuclear level.

The removal of all nuclear weapons from a zone of substantial width on both sides of the East-West boundary in central Europe would reduce the chances of the contingency most likely to precipitate such desperate decisions. It would remove these weapons from the chaos of warfare, at least in the early days. It would give leaders on both sides a most precious commodity: time. It would allow more careful and considered analysis of the tactical military situation and the possibility of persevering through the continued use of conventional forces only. It would reduce the concern on both sides that the other side would feel pressure to release its weapons hastily, and it would correspondingly lessen the urge to launch a preemptive attack. Finally, it would allow the greater exploration of diplomatic alternatives; during the increased time available before a decision to release nuclear weapons would have to be made, both sides might come to see the extraordinary dangers of the course on which they had embarked and take steps to resolve the conflict before it got out of hand.

There would be other advantages to implementing the commission's proposal for a battlefield-nuclear-weapon-free zone. The changes necessitated by such an agreement might lead both sides independently to reconsider the current inventories of nuclear weapons on the continent. Many observers have argued that substantial reductions are possible on military grounds. What has forestalled them are concerns about the political implications of withdrawing weapons unilaterally. The relocations made necessary by the battlefield-nuclear-weapon-free zone, and the fact that the other side was making similar changes, might provide the opportunity for substantially reducing nuclear forces without suffering adverse political consequences.

The successful negotiation of a battlefield-nuclear-weapon-free zone, along with the proposed measures to restrain conventional and chemical weapons, would open more ambitious possibilities for controlling and reducing military arsenals in Europe and elsewhere; it would have a positive effect on political relations between the Eastern and Western nations, implying the start of progress toward less tense and more cooperative relations between the U.S. and the U.S.S.R., and it would be good for relations within the Western alliance as well, eliminating what has become a damaging source of friction between the U.S. and Western Europe: the debate over approaches to nuclear-arms control on the Continent.

Late last year the Swedish government formally solicited the views of all European and North American governments on the proposed battlefield-nuclear-weapon-free zone. In response the U.S.S.R. stated in January that it endorsed the idea and would put such a proposal on the table at the Vienna talks. It also said, however, that the present proposal did not go far enough, adding that the zone should be between 500 and 600 kilometers wide. Such an expansion would take in virtually all of both East Germany and West Germany.

So far the Western nations have remained cool to the idea of a denuclearized zone. If the Russians do indeed put forward such a proposal in Vienna, however, there would be a golden opportunity to link the implementation of a denuclearized zone explicitly to the conclusion of a first-phase agreement on the limitation of conventional forces, and to negotiate a mutually acceptable width for the zone that did not give rise to political problems similar to those presented by the Rapacki plan.

The implementation of a battlefield-nuclear-weapon-free zone would not eliminate the risk of nuclear war in Europe. Some risk will remain as long as nuclear weapons remain in the inventories of the major powers. The degree of risk will wax or wane in consonance with overall political relations and specific changes in military deployments on the two sides, but it is unlikely ever to disappear. It is a fact with which everyone will have to continue to live. The proposed battlefield-nuclear-weapon-free zone, however, is a pragmatic and politically feasible measure to control the danger of nuclear war in Europe. It merits serious consideration by all citizens and governments.

Silicon Micromechanical Devices

Tiny valves, nozzles, pressure sensors and other mechanical systems can be chemically etched in a wafer of single-crystal silicon. Such devices can be mass-produced much as microelectronic circuits are

by James B. Angell, Stephen C. Terry and Phillip W. Barth

In the past 30 years the element silicon has become familiar as the material from which electronic components and systems are made. In the form of microelectronic chips it can be found in everything from dishwashers to the Space Shuttle control systems. The exploitation of silicon's electrical properties has been accompanied by a less publicized exploration of its other properties and their potential uses. This research has led to the development of a technology called micromachining that allows silicon to be made into mechanical devices almost as small as microelectronic ones.

Micromachining starts with the same batch-fabrication techniques that have made silicon integrated-circuit chips inexpensive. The techniques make it possible to fabricate many chips at once, so that the cost of production is spread over many chips and the cost per chip is low. To this repertory of techniques micromachining adds chemical etching techniques for forming three-dimensional shapes such as pits, holes, pyramids, trenches, hemispheres, cantilevers, diaphragms, needles and walls. A wide variety of mechanical devices can be constructed from combinations of these structural elements.

Among the silicon micromechanisms that have been built are valves, springs, mirrors, nozzles, connectors, printer heads, circuit boards, heat sinks and sensors for properties such as force, pressure, acceleration and chemical concentration. Even a device as complex as a gas chromatograph, an instrument for identifying and measuring gases in an unknown mixture, can be built on a disk of silicon a few centimeters in diameter.

Much of the interest in micromachining derives from the need for cheaper and more versatile sensors. Until recently the electronic components of automatic-control systems and measuring instruments were more expensive than the sensors. With the advent of the microprocessor the cost of electronics began to fall dramatically. Batch-fabricated silicon sensors could replace expensive hand-assembled sensors much as batch-fabricated silicon circuits have replaced vacuum-tube assemblies.

The latest development in this area is the integrated sensor, a silicon chip that includes both a sensor and associated signal-conditioning electronics. Integrated sensors are less expensive than sensors and the necessary electronic components fabricated separately. In addition, because the signal from an integrated sensor is less vulnerable to noise and leakage, integrated sensors perform better than discrete sensors and have a wider variety of applications.

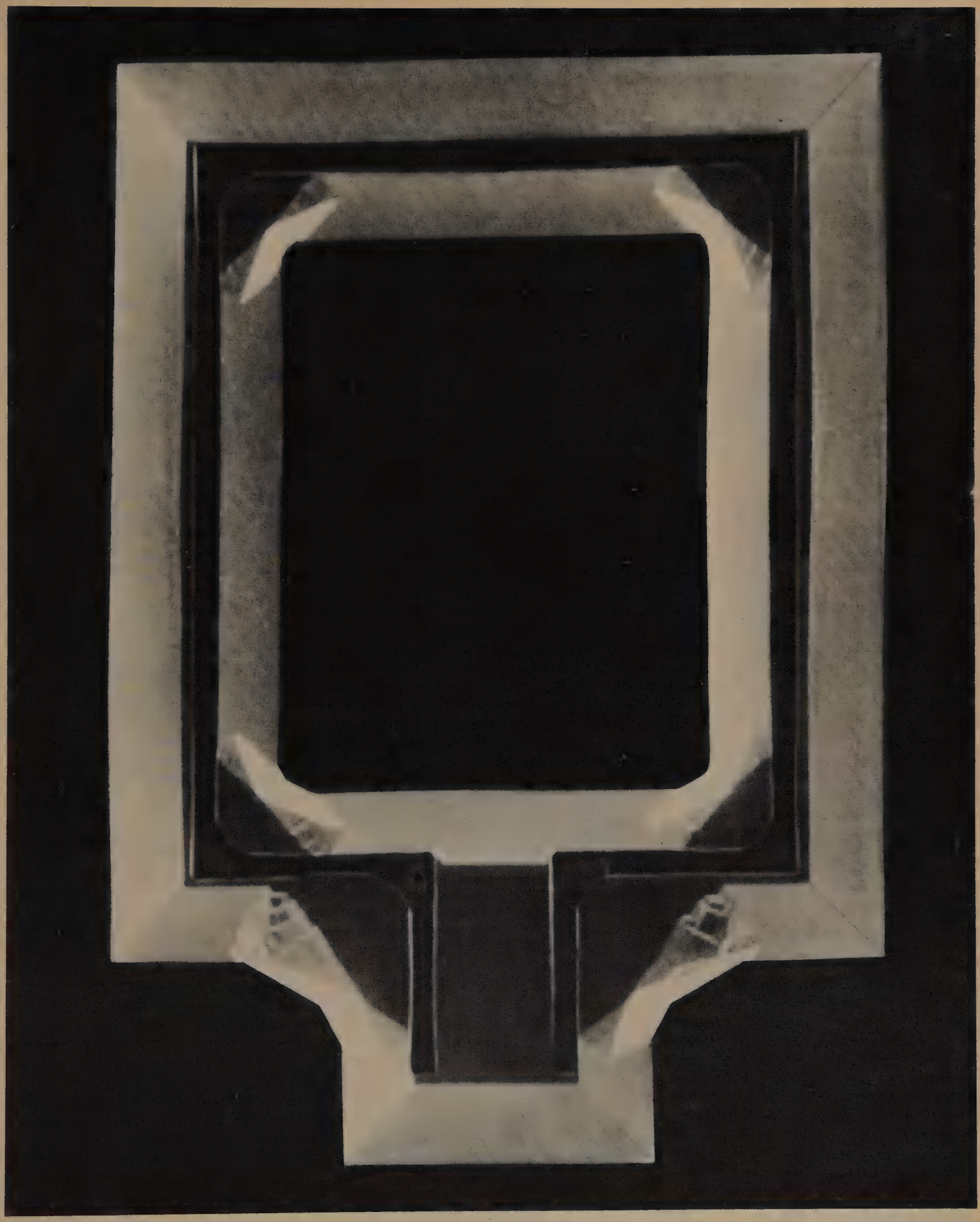
Silicon is a semiconductor, one of the elements that lie between the metals and the nonmetals in the periodic table of the elements. It is distinguished from other semiconductors, however, in that it can be readily oxidized. Silicon forms a surface layer of silicon dioxide (SiO_2) when it is exposed to steam. The layer is chemically inert and electrically insulating; essentially it is a glass. Oxide layers are used to protect areas of the silicon during the batch fabrication of microelectronic devices, which is why most such devices are made out of silicon rather than out of other semiconductors such as germanium. Micromachining also makes extensive use of silicon dioxide surface layers.

The silicon employed in the electronics industry takes the form of single crystals whose production is in itself a sophisticated technology. Large "boules," or single crystals of silicon, 10 centimeters in diameter and a meter long, are grown and sliced into disk-shaped wafers from .2 to .5 millimeter thick. The wafers are polished to a mirror finish to remove flaws created by sawing. The homogeneous crystal structure of the material gives it the electrical properties needed in microelectronic circuits. As it turns out, silicon in this form also has desirable mechanical properties.

Silicon, which is directly below carbon in the periodic table, forms the same type of crystal as diamond, although the interatomic bonds are somewhat weaker. Single-crystal silicon is brittle and can be cleaved like a diamond, but it is harder than most metals. In addition single-crystal silicon is surprisingly resistant to mechanical stress. In both tension and compression it has a higher elastic limit than steel; on the other hand, when the limit is reached, silicon fractures whereas steel deforms inelastically. Finally, as a single crystal silicon remains strong under repeated cycles of tension and compression, whereas polycrystalline metals tend to weaken and break because stresses accumulate at the intercrystal boundaries.

Both micromachining and microelectronic fabrication begin with photolithography, the photographic technique used to transfer copies of a master pattern onto the surface of a silicon wafer. The first step in photolithography is to grow a thin layer of oxide on the wafer surface by heating it to between 800 and 1,200 degrees Celsius in an atmosphere of steam. (Dry oxygen also works, but steam is much faster.) Next, a thin layer of an organic polymer that is sensitive to ultraviolet radiation, called a photoresist, is deposited on the oxide surface. A photomask, generally a glass plate on which there is a metal pattern, is placed in contact with the photoresist-coated surface, and the wafer is exposed to ultraviolet radiation. The metal on the mask is opaque to ultraviolet wavelengths, whereas the glass is transparent. The radiation causes a chemical reaction in the exposed areas of the photoresist. If the photoresist is of the type called positive, the chemical reaction weakens the polymer. If the photoresist is of the type called negative, the reaction strengthens the polymer. The wafer is rinsed in a developing solution that removes either the exposed areas or the unexposed areas of photoresist, leaving a pattern of bare and photoresist-coated oxide on the wafer surface. The photoresist pattern is either the positive or the negative image of the pattern on the photomask.

The wafer is then placed in a solution



SILICON ACCELEROMETER developed by Lynn M. Roylance of Stanford University is a miniature instrument that exploits both the mechanical and the electrical properties of silicon. It is made by selective chemical etching of a "wafer" cut from a large single crystal of silicon. The accelerometer, shown here in a scanning electron micrograph, is essentially a mass of silicon with a thickness equal to that of the wafer, suspended at the end of a thin silicon beam. An electrical resistor is fabricated in the underside of the beam; its resistance changes when the beam flexes, yielding a measure of acceleration. The mass is freed from the surrounding material by etching completely through the silicon at the periphery of the mass. The beam is defined by etching away all but a thin layer of silicon in that part of the wafer. The packaged device covers an area of two millimeters by three millimeters and is .6 millimeter thick, small enough to be sutured to the heart to measure its acceleration. Hundreds of accelerometers can be fabricated simultaneously on one silicon wafer.

of hydrofluoric acid, which attacks the oxide but not the photoresist or the underlying silicon. The photoresist protects the oxide areas it covers. Once the exposed oxide has been etched away the remaining photoresist can be stripped off with hot sulfuric acid, which attacks the photoresist but not the oxide or the silicon. The final result is a pattern of oxide on the wafer surface that duplicates the photoresist pattern and is therefore either a positive or a negative copy of the pattern on the photomask. The oxide pattern itself serves as a mask in subsequent processing steps.

At this point micromachining diverges from microelectronic fabrication. In processing an electronic device the oxide pattern serves as a mask during the "doping" of the wafer with impurities such as boron or phosphorus. When the wafer is heated, the impurity atoms deposited on its surface diffuse a short distance into the silicon underlying the openings in the oxide, creating the shallow conductive or resistive regions of which microelectronic devices are composed. In micromachining the oxide is used as a mask during chemical etching. The etchants attack the silicon underlying openings in the oxide layer, excavating deep three-dimensional pits in the wafer.

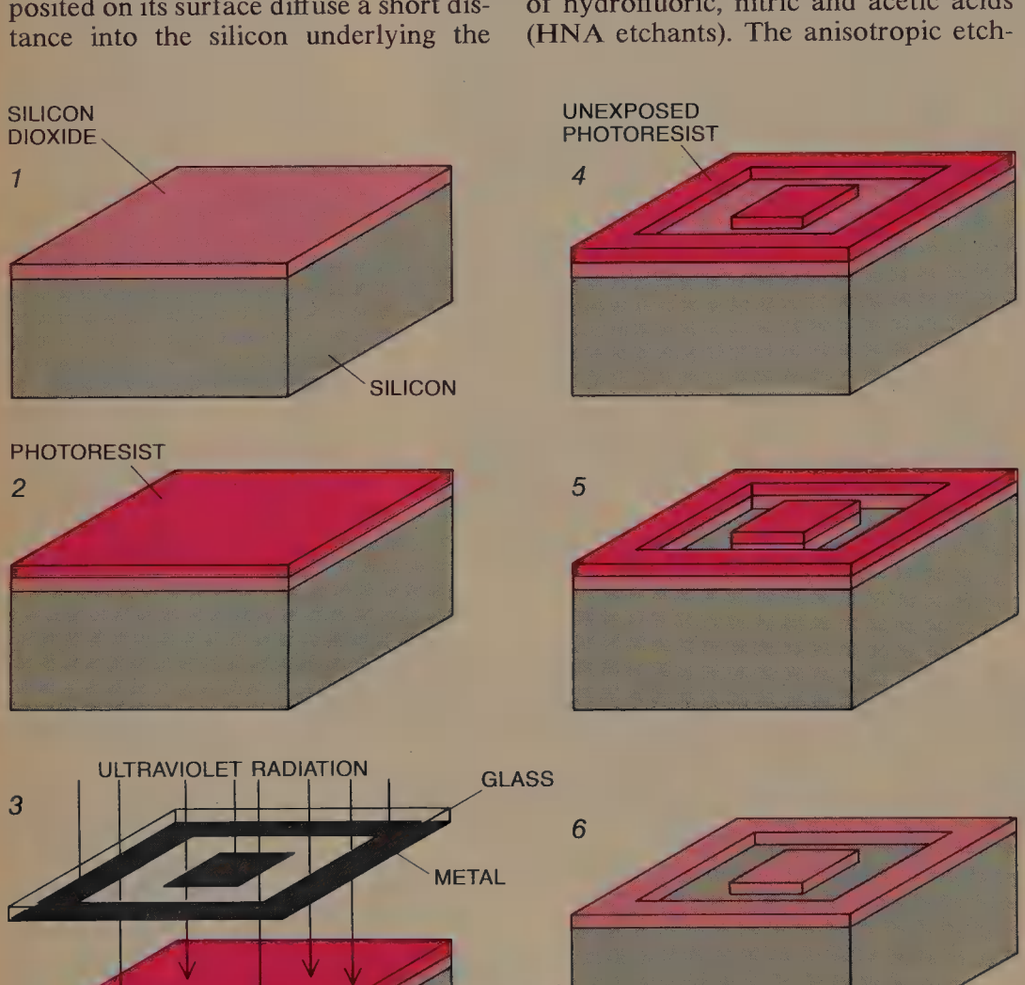
Two types of etchants are employed in the micromachining of silicon. Isotropic etchants etch the silicon crystal at the same rate in all directions and create gently rounded shapes. Anisotropic etchants, which are also known as orientation-dependent or crystallographic etchants, etch at different rates in different directions in the crystal lattice; they can form well-defined shapes with sharp edges and corners. The most useful isotropic etchants are mixtures of hydrofluoric, nitric and acetic acids (HNA etchants). The anisotropic etchants are all hot alkaline solutions such as aqueous potassium hydroxide (KOH), aqueous sodium hydroxide (NaOH) and a mixture of ethylenediamine, pyrocatechol and water known as EDP.

Micromachining makes more extensive use of the anisotropic etchants than of the isotropic ones. A typical micromachining operation, the etching of an array of holes completely through the wafer, illustrates why. An isotropic etchant moves both downward and outward from an opening in the oxide, undercutting the oxide mask and enlarging the etched pit while deepening it. The pit created is at least twice as wide as it is deep, so that the openings in the oxide must be placed at least as far apart (edge to edge) as the wafer is thick if separate holes are to be etched through the wafer. On the other hand, an anisotropic etchant, if it is properly employed, does not undercut an oxide mask and creates pits with well-defined side walls. The pits deepen without widening. Thus an anisotropic etchant can create closely spaced arrays of holes; the edges of the openings in the oxide mask can be placed as close together as the limits of photolithography allow.

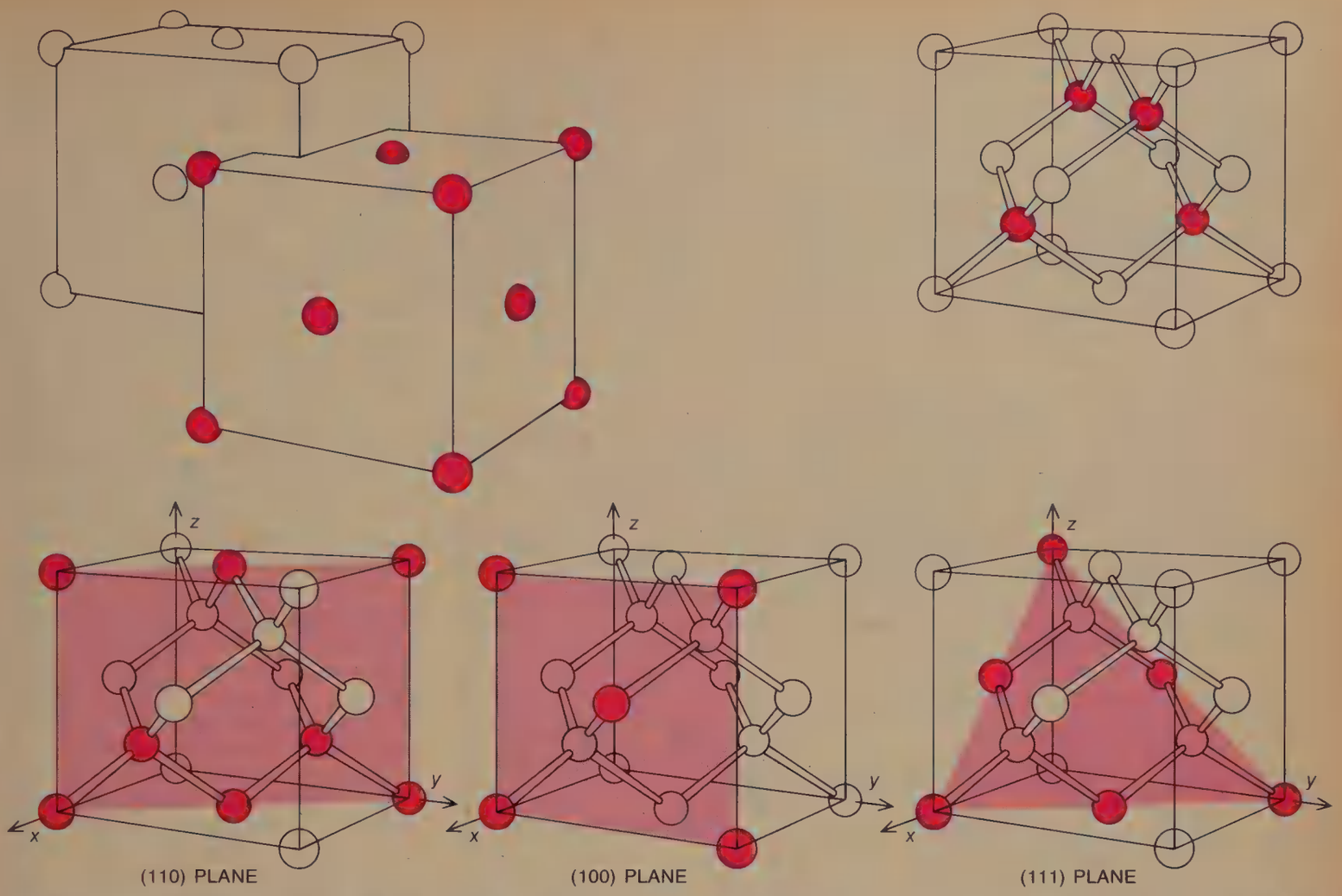
The shape of a hole formed by an anisotropic etchant depends on the orientation of the atomic planes in the silicon wafer. Just as various one-dimensional rows of corn stalks can be identified at different angles in a two-dimensional cornfield, so various two-dimensional planes of atoms lie at different angles in a three-dimensional crystal. The planes of atoms in the crystal are identified by means of a coordinate system assigned to the crystal.

Each silicon atom bonds with four neighboring atoms. The bonding structure is tetrahedral, that is, each atom lies at the center of a tetrahedron defined by the four atoms with which it is bonded. The bonding structure is an inconvenient basis for a coordinate system because the atomic bonds are not at right angles to one another. It is more convenient to think of the silicon crystal as being composed of stacked layers of repeating cubes. Each cube has an atom at each corner and at the center of each face and interlocks with four neighboring cubes. This interlocking face-centered cubic structure can serve as the basis of a coordinate system. Any atom can be designated as the point of origin of a coordinate system whose axes lie along the edges of the cube to which the atom belongs. The basic unit of measure is the length of the edge of one cube.

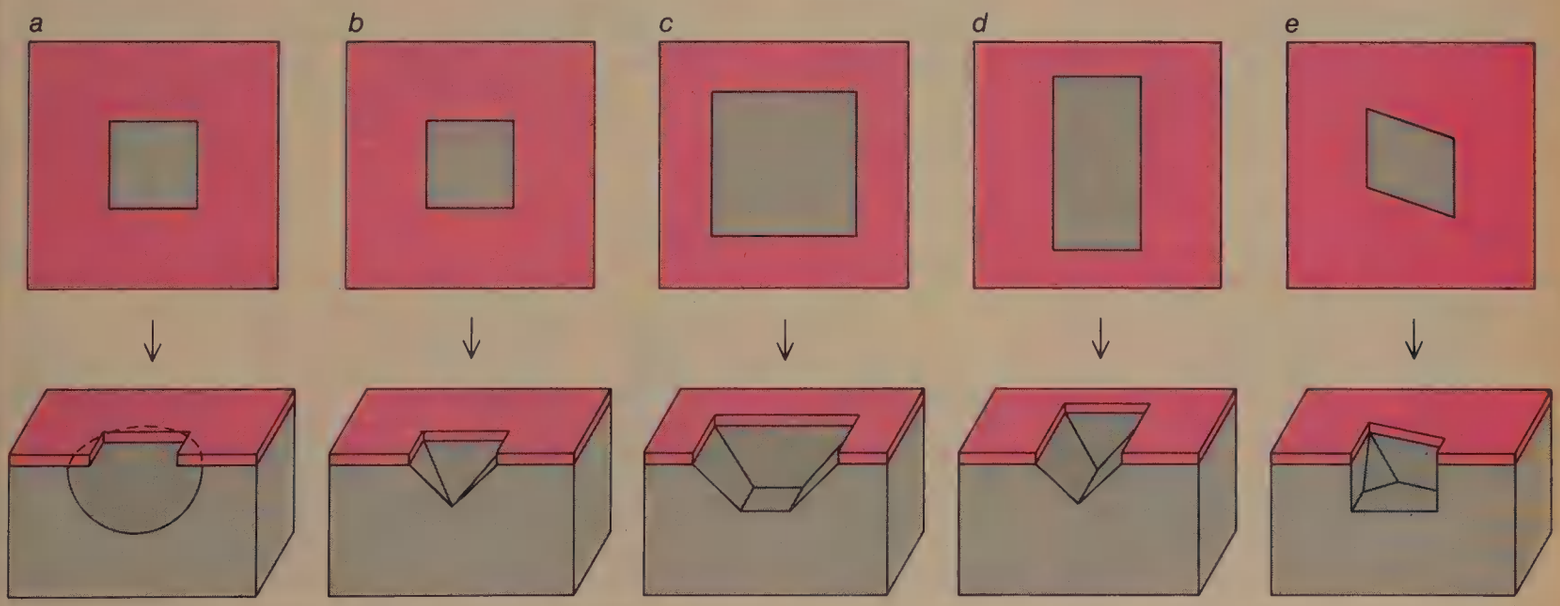
A vector, or direction in the crystal, is described by three coordinates known as the Miller indexes. For example, a [110] vector points diagonally across the face of a unit cube. The coordinates of a given vector also designate the entire set of atomic planes perpendicular to that direction. Thus the notation (100) de-



OXIDE PATTERNS on the wafer surface protect selected areas during the wet chemical etching that shapes a micromechanical device. First, a layer of silicon dioxide (SiO_2) is grown on the wafer surface; silicon reacts with steam at temperatures between 800 and 1,200 degrees Celsius to form the oxide. The oxidized wafer is coated with a polymer sensitive to ultraviolet radiation, called a photoresist. Next the wafer is placed under a master mask (a metal pattern on a glass plate) through which it is exposed to radiation. Photoresist in exposed areas weakens and is rinsed away in a developing solution. Then the wafer is etched in an acid that attacks the uncovered oxide but not the remaining photoresist or the silicon. The result is a pattern of openings in the oxide that duplicates the metal pattern on the glass plate. In subsequent processing steps the oxide pattern serves as a mask for the etching of the underlying silicon.



CRYSTALLINE SILICON shares with diamond the crystal structure called interlocking face-centered cubic. An atom lies at each corner and in the center of each face of the unit cube, and the cubes interlock in such a way that several atoms from neighboring cubes lie within each unit cube. The axes of the unit cube provide a rectilinear coordinate system that makes it possible to specify directions and planes within the crystal. A crystalline direction is designated by three coordinates called the Miller indexes, which are integer multiples of the length of one edge of a unit cube. The same set of indexes designates the planes perpendicular to the direction. Crystalline orientation is important in the fabrication of micromechanical devices because some of the etchants employed attack different directions in the crystal at different rates. Most such anisotropic etchants progress rapidly in the crystal direction perpendicular to the (110) plane and less rapidly in the direction perpendicular to the (111) plane. The direction perpendicular to the (111) plane etches very slowly, if at all.



ANISOTROPIC ETCHANTS create faceted holes composed of the crystal planes that are etched slowest, unlike isotropic etchants, which invariably produce a gently rounded hole (*a*). The shape of an anisotropically etched hole is determined by the crystalline orientation of the wafer surface, the shape and orientation of the openings in the mask on that surface and the orientation dependence of the etchant itself. A square opening oriented along the $<110>$ directions of a $<100>$ wafer yields a pyramidal pit with $\{111\}$ side walls (*b*). If the opening is larger, the point of intersection of the $\{111\}$ planes is deeper, and a flat-bottomed pit can be created by stopping the etch before that depth is reached (*c*). A rectangular opening on the same wafer gives a *V*-shaped groove (*d*). Holes with parallel side walls can be created by etching a wafer that has a different surface orientation. In a $<110>$ wafer two sets of $\{111\}$ planes are perpendicular to the surface although not to each other. If an oxide opening on a $<110>$ wafer is properly oriented, etching creates a hole with vertical side walls. The side walls that intersect at acute angles are bridged by still other planes. These are just a few of the many shapes that can be etched.

scribes all the planes perpendicular to the x axis. Because the crystal structure is symmetrical, the x, y and z directions are interchangeable, and a generalized notation can be introduced to describe all equivalent directions and sets of planes. The notation $<110>$ designates the diagonals across any face of a unit cube, and the notation $\{110\}$ designates the sets of planes perpendicular to all $<110>$ vectors. The different brackets and parentheses distinguish planes from directions and distinguish generalized planes and directions from particular planes and directions.

The mechanisms that underlie the dependence of the etch rate of an anisotropic etchant on crystal direction are not well understood. The differences in etch rate depend on temperature and on the composition of the etchant, and they also seem to be related both to the density of atomic bonds on an exposed silicon plane and to the radius of the hydrated hydroxyl ions in the etching solution.

Experience has shown, however, that exposed $\{111\}$ planes are etched slowly, if at all, by the anisotropic etchants. In contrast, $\{110\}$ planes are usually etched rapidly, and $\{100\}$ planes are etched at an intermediate rate. Simple etched holes may have only $\{111\}$ and $\{100\}$ facets. Other higher-order planes, such as $\{221\}$ planes, are also etched at an intermediate rate and may appear as facets of complex etched shapes.

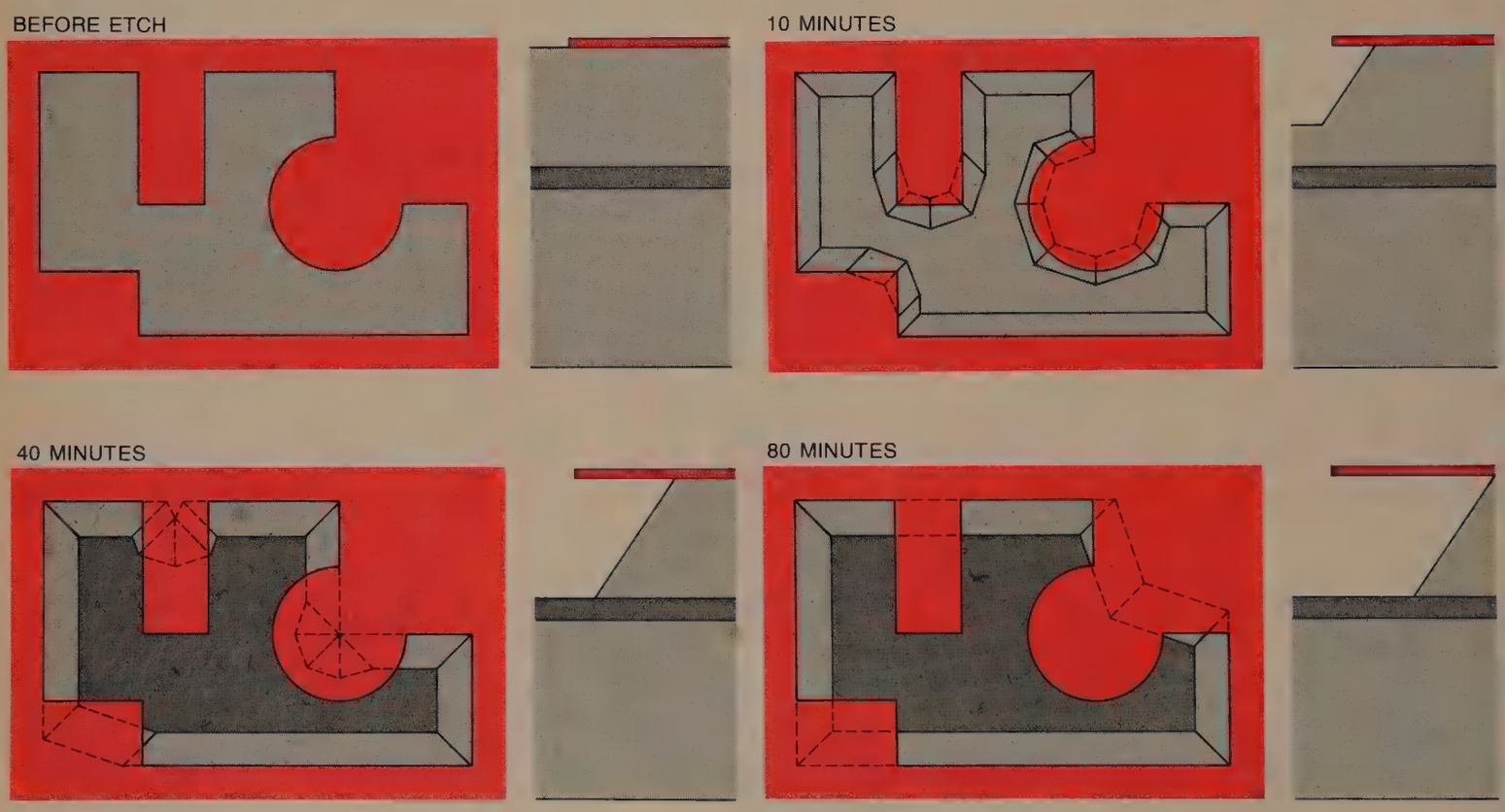
The type of hole an anisotropic etchant forms in a wafer is determined by the crystal orientation of the wafer surface as well as by the orientation dependence of the etchant. The two types of silicon wafers most useful for anisotropic etching are $<100>$ wafers (wafers in which $\{100\}$ planes are parallel to the surface and a $<100>$ direction is perpendicular to it) and $<110>$ wafers (in which $\{110\}$ planes are parallel to the surface and a $<110>$ direction is perpendicular to it).

The orientation, size and shape of the oxide opening on the wafer surface also play a part in determining the type of hole formed, as an example illustrates. If a square hole is opened in an oxide layer on a $<100>$ wafer, treatment with anisotropic etchant can create an upside-down pyramidal pit. Four sets of $\{111\}$ planes intersect the surface of a $<100>$ wafer along the two perpendicular $<110>$ directions that lie in the surface plane. During anisotropic etching the exposed $\{100\}$ surface plane is etched downward at a constant rate, giving the pit a flat bottom at the start of the etch process. At the edges of the hole four inward-slanting $\{111\}$ planes (at angles of 55 degrees to the surface) form the walls of the pit. As etching continues, more of the $\{111\}$ planes is exposed and the area of the flat bottom shrinks. Eventually the $\{111\}$ planes meet at a point and the flat bottom is gone. The etching then stops because no readily attacked planes are exposed.

The square oxide opening must be precisely aligned (within one or two degrees) with the $<110>$ directions on the wafer surface to obtain pyramidal pits that conform exactly to the oxide mask rather than undercutting it. Most silicon wafers have a flat edge parallel to a $<110>$ direction in the wafer. The square mask pattern is lined up with this flat edge during the photolithography step that precedes the etching.

The size of the oxide pattern determines not only the area of the pit on the wafer surface but also the depth of the pit. The larger the square oxide opening, the deeper the point at which the $\{111\}$ side walls of the pit intersect. If the oxide opening is large enough, the $\{111\}$ planes do not intersect within the wafer. The etched pit therefore extends all the way through the wafer, creating a small square opening on the bottom surface.

The influence of the shape of the oxide opening on the geometry of the etched pit is straightforward. If the opening in the oxide on a $<100>$ wafer



OXIDE CANTILEVER can be formed by undercutting an oxide etching mask. Convex oxide shapes are undercut by an anisotropic etchant, although properly oriented concave shapes are not. The rectangular oxide opening shown here is a concave shape; the oxide tabs protruding into the opening are convex shapes. If the downward progress of the etchant is limited by an etch-stop layer, a prolonged anisotropic etch ultimately yields an oxide tab overhanging a shallow rectangular pit with $\{111\}$ side walls and a flat bottom, namely the etch-stop layer. Cantilevers like the one formed when the oxide shape at the upper left is completely undercut have been used to make accelerometers, arrays of electrostatically controlled switches and arrays of mirrors. Undercutting makes it difficult to etch a hole into which a right-angle silicon corner projects. The silicon under the right-angle oxide shape at the lower left is rounded off during etching. A roughly right-angle silicon corner is formed under the circular oxide shape at the upper right at one point during the etch. The corner is eroded as the etch continues, however. Timed etches of patterns with circular tabs can be employed to make holes with projecting right-angle corners.

is rectangular rather than square, etching creates a long trench rather than an inverted pyramid. The walls and the ends of the trench are {111} planes, and if the etch goes to completion, the trench has a *V*-shaped cross section.

Pyramidal pits, *V*-shaped grooves and other faceted pits and holes are one kind of structural element of which micromechanical devices are composed. A second structural unit is a thin membrane of silicon. A membrane can be formed by etching a wafer for a period just short of what it would take to etch through it. It is difficult to make diaphragms of uniform thickness by this method, however, because the thickness of the diaphragm is determined by that of the wafer, and wafer thickness typically varies by 10 micrometers or more both across a single wafer and from one wafer to another.

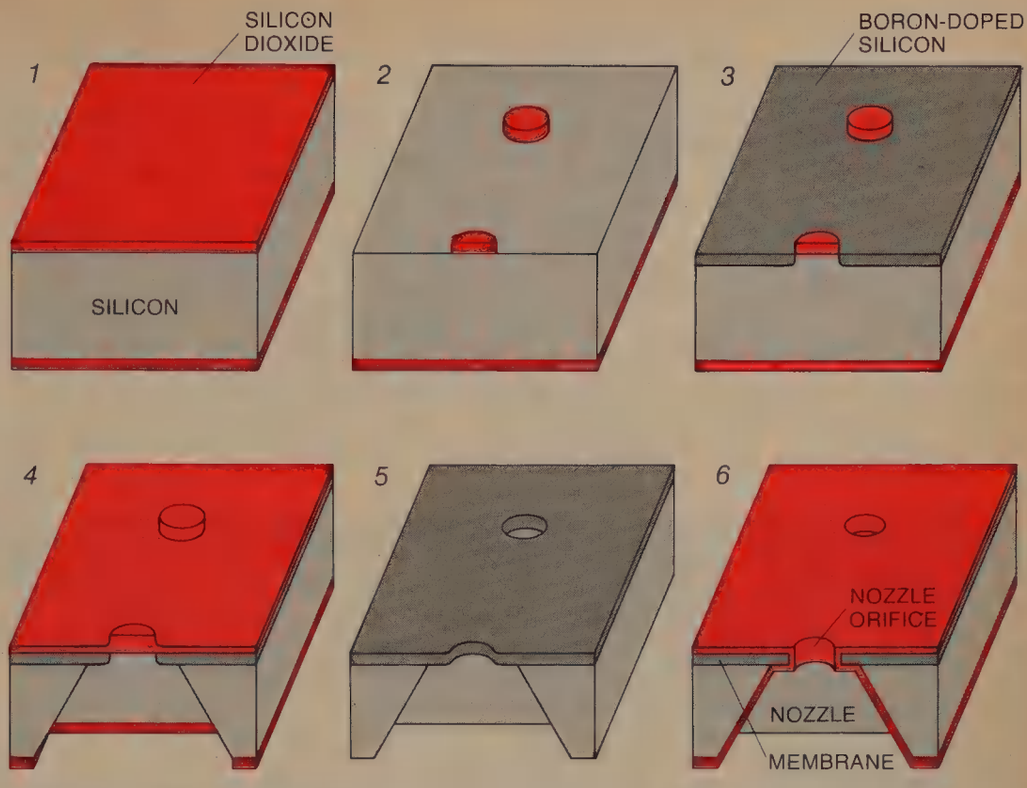
Membranes of more precisely defined thickness and shape can be created by a technique that exploits another property of anisotropic etchants. The rate at which the solutions etch a wafer depends not only on crystal orientation but also on the extent to which the silicon has been doped with impurity atoms such as boron, phosphorus, arsenic or antimony.

The etch rate of a doping-dependent etchant (a category that overlaps the categories of isotropic and anisotropic etchants) depends on the type of dopant atoms and their concentration. The isotropic etchant HNA is doping-dependent in some mixture ratios; it etches heavily doped silicon much faster than it etches lightly doped silicon. On the other hand, the anisotropic etchants potassium hydroxide and EDP etch silicon that has been heavily doped with boron much more slowly than they etch silicon lightly doped with boron.

The first step in creating a membrane by anisotropic etching is to create an etch-stop layer of silicon heavily doped with boron on the surface of the wafer. A layer of some substance rich in boron is deposited on the front surface of the wafer, which is then heated in a furnace to between 1,000 and 1,200 degrees C. At these temperatures the boron atoms diffuse into the silicon. The depth of the doped layer is controlled by the temperature and the duration of the diffusion step. If the wafer is then etched from the back with an anisotropic etchant, the etchant eats through to the doped layer and stops, creating a membrane as thick as the doped layer on the surface of the wafer.

If the designer wants to put a membrane some distance below the wafer surface, he dopes the surface in much the same way and then grows a layer of silicon over the doped layer by heating the wafer in an atmosphere of silane gas (SiH_4). The gas decomposes at high temperature and deposits silicon on the wafer surface. The new layer, which is called an epitaxial layer, assumes the crystal structure and orientation of the silicon on which it is deposited. The epitaxial layer is typically from five to 20 micrometers thick.

If the wafer is then etched from both sides in an anisotropic etchant, a membrane is created at a depth equal to the thickness of the epitaxial layer. The thickness of the membrane itself is again determined by the thickness of the boron-doped layer.



SILICON INK NOZZLE in one design for a high-speed ink-jet printer is a pyramidal pit whose bottom is a thin membrane through which an orifice of precise diameter has been etched. The diameter of the orifice is determined by the first three processing steps. The wafer is oxidized and dots of photoresist are defined on the top surface. The oxide not protected by photoresist is etched away, leaving a pattern of oxide dots. When boron is diffused into the silicon, it creates a layer of boron-doped silicon everywhere except under the dots. Then a square opening is defined in the oxide on the bottom of the wafer and is etched with an anisotropic etchant. The etchant creates a pyramidal pit with {111} side walls but stops when it comes to the doped layer. It eats through the undoped dots in the layer, however, so that when all the oxide is removed, the nozzle orifices penetrate the wafer. Finally, the wafer is reoxidized to give all surfaces a protective coating. Several arrays of nozzles are fabricated simultaneously on one wafer.

A third example of a micromechanical structure is the cantilever beam, a thin beam of silicon dioxide supported at only one end. In many applications a cantilever is suspended over a shallow pit in the silicon. The techniques for making a cantilever and a pit were developed by a group working with Kurt E. Petersen at the San Jose Research Laboratory of the International Business Machines Corporation. The techniques exploit two properties of anisotropic etchants: their doping dependence and their tendency to undercut convex shapes in an oxide mask.

The simple geometric oxide openings discussed so far are concave oxide shapes, that is, the line between the surrounding oxide and the opening is always concave with respect to the opening. A concave oxide shape is not undercut by an anisotropic etchant if the opening is properly oriented on the wafer surface. If a finger or a tab of oxide projects from one side of such an opening, however, its boundary is convex with respect to the exposed silicon. Such convex oxide shapes are undercut by an anisotropic etchant.

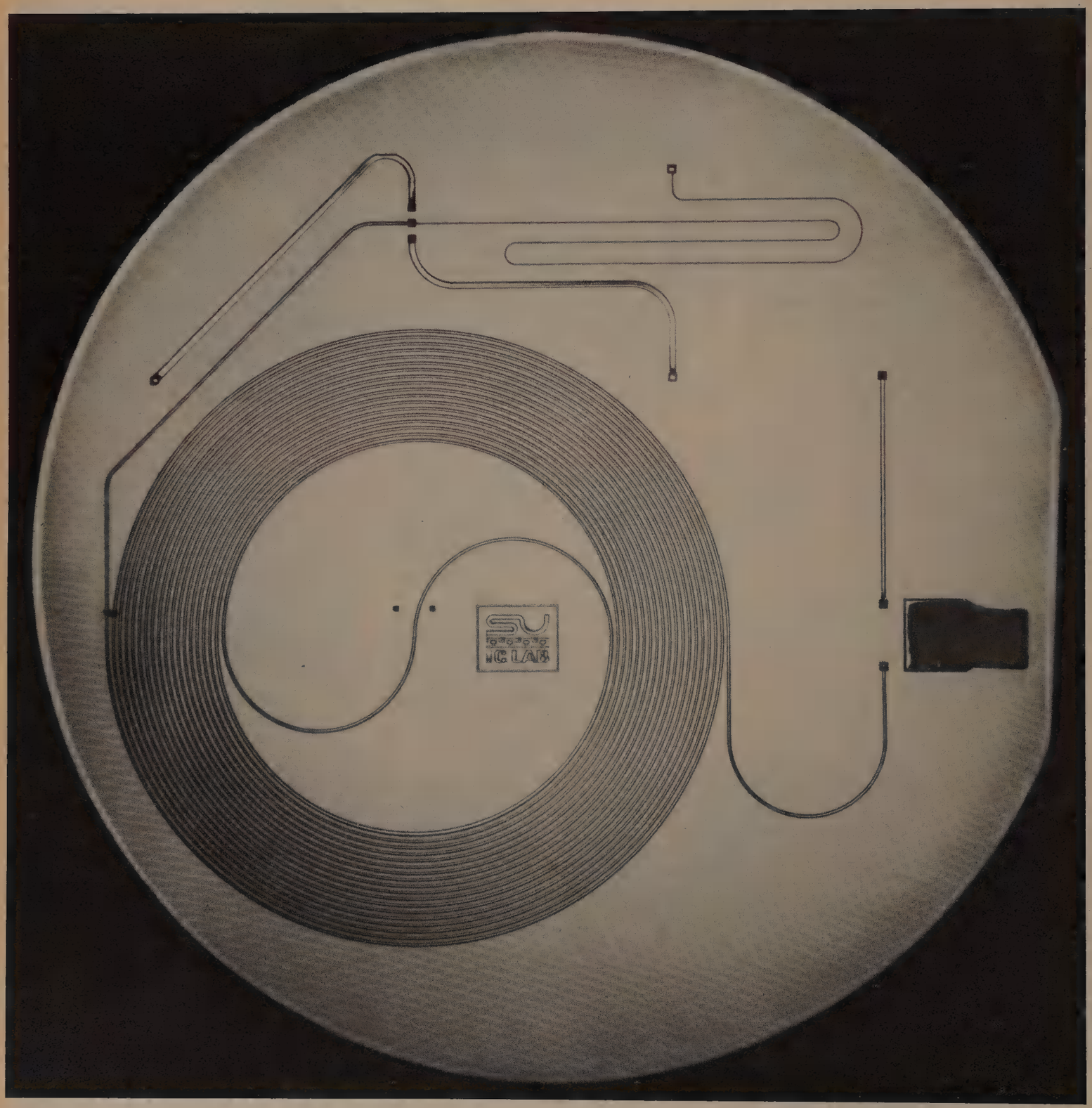
The first step in making a cantilever over a shallow pit is the formation of a boron etch-stop layer. A layer of epitaxial silicon is laid down over the boron-doped stratum, oxide is grown over the epitaxial silicon and an opening is made in the oxide. The wafer is then etched in an anisotropic etchant. If an oxide tab protrudes into the oxide opening, the etchant begins to undercut it. At the attached end of the oxide tab the etchant eventually runs into a {111} plane and stops etching. Since the etch-stop layer limits the downward progress of the etch, the final result is a cantilever of oxide extending over a shallow depression in the substrate.

In the Integrated Circuits Laboratory at Stanford University work in micromachining began in 1965 when Kensall D. Wise (now at the University of Michigan) and one of us (Angell) developed a closely spaced array of elec-

trodes for recording electric potentials in the brain. The project led to the development of a variety of brain probes. Since then projects in micromachining at Stanford have included pressure sensors, accelerometers, gas chromatographs, miniature thermometer arrays and structures for the dielectric isolation of microelectronic circuits. Micromechanical techniques and devices are also being developed at several other universities and industrial laboratories. The devices we shall describe below include some of our own and some developed by others.

Arrays of nozzles fabricated in silicon have found application in ink-jet printers for digital computers. In one type of ink-jet printer ink is forced through a linear array of nozzles, creating several parallel streams that are broken into droplets as they leave the nozzles. As droplets in each stream pass single file through a charging electrode, each droplet is either given an electric charge or allowed to remain electrically neutral. An uncharged droplet passes undeflected through an electrostatic field and leaves a dot on the paper, but a charged droplet is deflected by the field and instead of hitting the paper is intercepted by a collection gutter. A printed character is formed by selecting droplets from the vertical array of nozzles as the entire mechanism moves horizontally across the paper.

Ernest Bassous, Larry Kuhn and their colleagues at the IBM Thomas J. Watson Research Center did the pioneering work in micromechanical ink-jet nozzle arrays, developing many different nozzle systems. The simplest way to make a nozzle is to etch a pyramidal pit in the silicon. If the intersection point of the {111} planes that form the walls of the



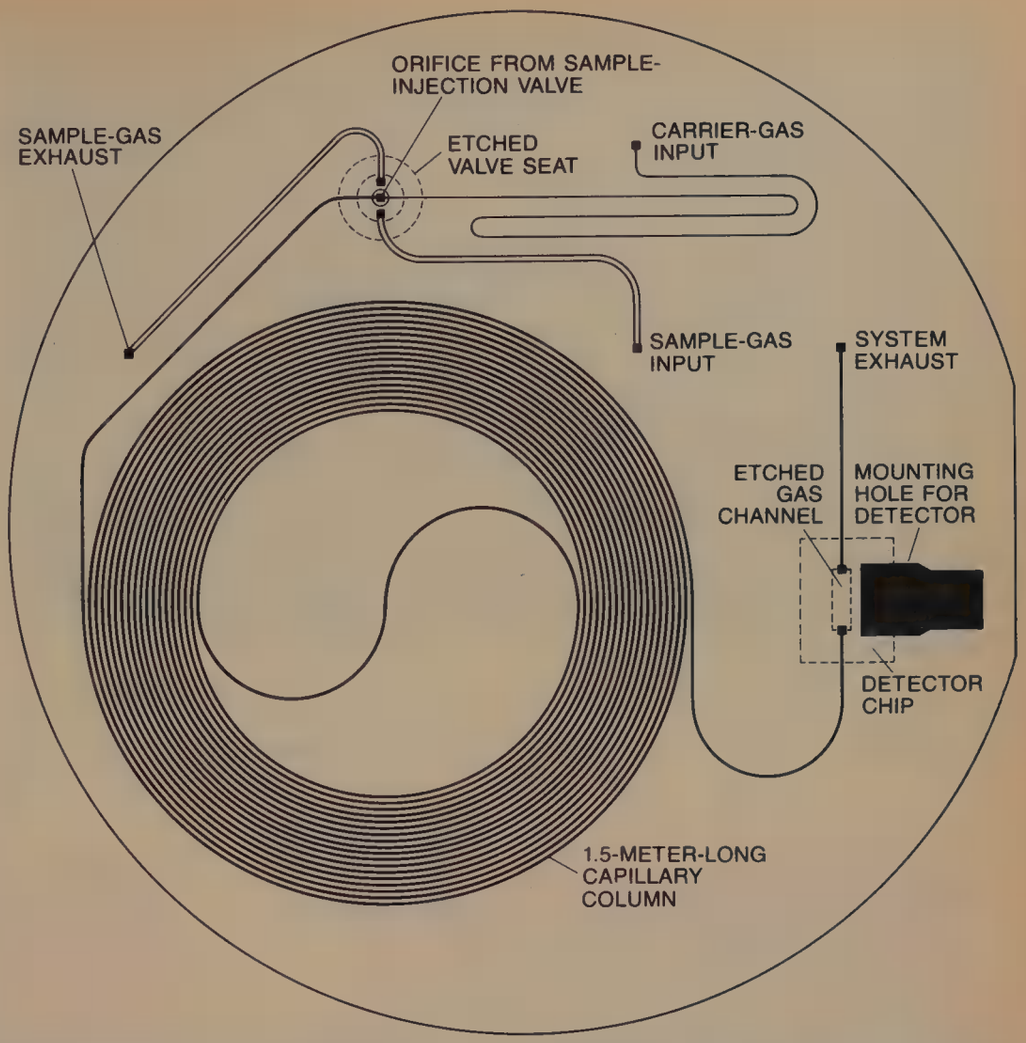
SILICON GAS CHROMATOGRAPH is a complete analytic instrument built on a wafer five centimeters across. Capillary column of the chromatograph is a spiral 1.5 meters long. The column is formed by etching a groove in a wafer and bonding the wafer to glass. This prototype was developed at Stanford University with funds provided by the National Institute for Occupational Safety and Health.

pit is below the bottom surface of the wafer, a small square opening is created. It is difficult to make nozzle orifices of uniform size with this technique, however, in part because the dimensions of the orifice depend on the thickness of the wafer; as noted above, the thickness is hard to control.

One nozzle design solves the problem by putting a membrane three micrometers thick at the bottom of each nozzle cavity. A pyramidal pit, which serves as the cavity, is etched down to the membrane, but the nozzle orifice is defined by a boron-diffusion procedure, not by the geometry of the pit. By this method Bassous and his co-workers have built arrays of eight nozzles with circular orifices exactly 20 micrometers in diameter spaced exactly .3 millimeter apart.

The first step in making the array is to define the nozzle orifices as dots of photoresist on the front surface of a wafer coated with oxide on both sides. All the oxide on the front surface is etched away except for the dots protected by photoresist. Then the wafer surface is doped with boron, which is driven into the silicon to a depth of three micrometers by heating the wafers in a diffusion furnace. Only the silicon under the oxide dots remains undoped.

Next, photolithography and an etching step are used to define a pattern of square openings in the oxide on the back surface of the wafer. The openings are centered under the undoped dots on the front surface. The wafer is etched in a doping-dependent anisotropic etchant, which eats through 99 percent of the wafer's thickness but stops when it comes to the heavily doped boron layer. It etches all the way through the undoped dots, however, creating the nozzle orifices. The diameter of each orifice, the thickness of the surrounding membrane and the placement of the orifices with respect to one another can be controlled with micrometer accuracy because they are defined not by the wafer thickness or the length of the etching step but by photolithography and the boron-diffusion step.



SEPARATION OF GASES in the silicon gas chromatograph is based on differences in the solubility of various gases in a liquid that lines the capillary column. An inert carrier gas flows continuously through the capillary-column channel. When a valve in the channel is opened, a pulse of the gas to be analyzed is fed into the column and flushed through it by the carrier gas. As the gases in the sample pass through the column they are repeatedly adsorbed and desorbed on a thin liquid lining. Each gas is identified by its retention time in the column. As each gas arrives in turn at the end of the column, it passes through a hole to a channel on the back of the wafer over which a thermal-conductivity detector is mounted. Sample gases have a lower thermal conductivity than the helium carrier gas has and cause voltage peaks in the detector output. The volume of each gas is determined from the area under the voltage peak it creates.

The most elaborate of the micromechanical devices made so far is a gas chromatograph the size of a box of kitchen matches. Two of us (Terry and Angell) built a prototype of the device in 1975. It is now being developed as a commercial product by Microsensor Technology, Inc.

Gas chromatography is an analytic technique for separating, identifying and measuring the quantity of each gas in a mixture. A sample of the mixture is injected by a valve into a long capillary column, through which it is flushed by an inert carrier gas, in most cases helium. The walls of the capillary column are lined with a thin layer of a material such as a silicone oil or a polymer in which different gases have different degrees of solubility. As the component gases pass through the column they are repeatedly adsorbed and desorbed in the lining. Because the time a component gas remains adsorbed depends on its solubility, each gas travels through the column at a different rate and emerges at a specific time. The output stream is passed over a detector that measures a property of the gas such as its thermal conductivity. The output signal generated by the detector is a series of bell-shaped peaks corresponding to the sample gases, separated by flat regions corresponding to the inert carrier gas. Many modern gas chromatographs have an associated microprocessor that identifies each gas by matching its retention time with known retention times and measures the quantity of each gas by calculating the area under the output peak.

Like a conventional gas chromatograph, the miniature one consists of a capillary column, a sample-injection valve and a thermal-conductivity detector. All the parts of the miniature gas chromatograph, however, fit on a single wafer five centimeters in diameter. The volume of the capillary column is much smaller than that of a conventional capillary column. Because a chromatograph operates properly only if the volume of the injected sample gas is much smaller than the volume of the column, it was necessary to design a miniature sample-injection valve to accompany the miniature capillary column. Because the internal volume between the valve and the column and between the column and the detector must also be minimized, the valve and column are fabricated on the same wafer and the detector is a silicon chip mounted on the wafer.

The capillary column is a groove 1.5 meters long, wound into a spiral so that it fits on the wafer. A glass plate bonded to the wafer forms the top surface of the column. At the input end of the column a hole leads to the bottom surface of the

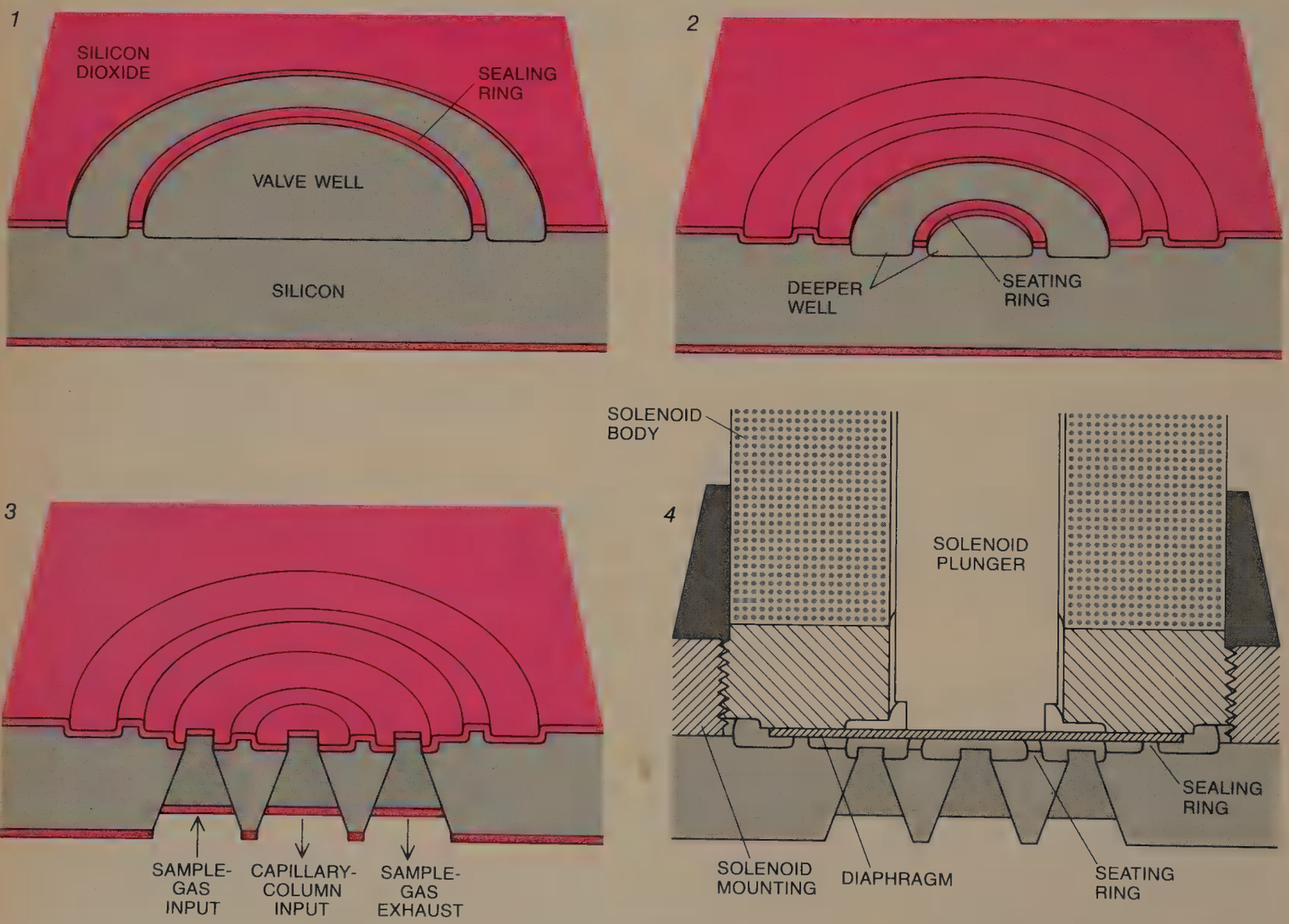
wafer. The helium carrier gas enters the column through this hole. A short distance away another hole in the column channel leads to a valve on the back surface of the wafer. The sample gas enters a separate channel through yet another hole, flows through the channel to the valve and is injected by the valve into the capillary column.

The valve seat on the back surface of the wafer consists of a silicon sealing ring, which surrounds both the input and the output orifices, and a silicon seating ring, which surrounds only the output orifice leading to the capillary column. A diaphragm of nickel and Teflon is clamped against the sealing ring. Normally a spring holds the plunger of a solenoid against the diaphragm, pushing it against the seating ring. When the solenoid is actuated, the plunger withdraws and the diaphragm relaxes, allowing gas to flow from the input orifice over the seating ring and into the output orifice. The effective dead volume of the valve is the volume of the capillary-column orifice, which is four nanoliters.

At the output end of the capillary column another etched hole leads to a gas channel etched into the bottom surface of the wafer. The chip on which the thermal-conductivity detector is built is inverted over this channel and clamped to the wafer. The detector is a thin-film metal resistor on a thermally isolating membrane of Pyrex glass in the middle of the chip. A constant electric current is passed through the resistor, and its resistance is monitored. Sample gases have a thermal conductivity lower than that of the carrier gas and remove less heat from the resistor, increasing its resistance and creating voltage peaks. The amplitude of each voltage peak is proportional to the quantity of one gas in the mixture.

The micromachining of the chromatograph wafer begins with the etching of the valve seat. The valve well, the sealing ring and the seating ring are defined by isotropic etches through concentric circular openings in the oxide layer on the back of the wafer. Then the feedthrough holes in the valve and at the ends of the capillary column and the sample-gas channels are formed by an anisotropic etch through square openings defined in the oxide on the front surface of the wafer. Finally the capillary column and the carrier-gas channels are defined by an isotropic etch through an oxide pattern on the front surface.

At this point the capillary column is a shallow spiral groove about 200 micrometers across at the surface of the wafer and 40 micrometers deep. The open spiral is made into an enclosed channel by stripping the oxide off the surface and bonding the wafer to a plate of Pyrex glass. The bonding technique, called anodic bonding, requires no intermediate liquid layer such as a glue or solder. A method of this kind is essential since



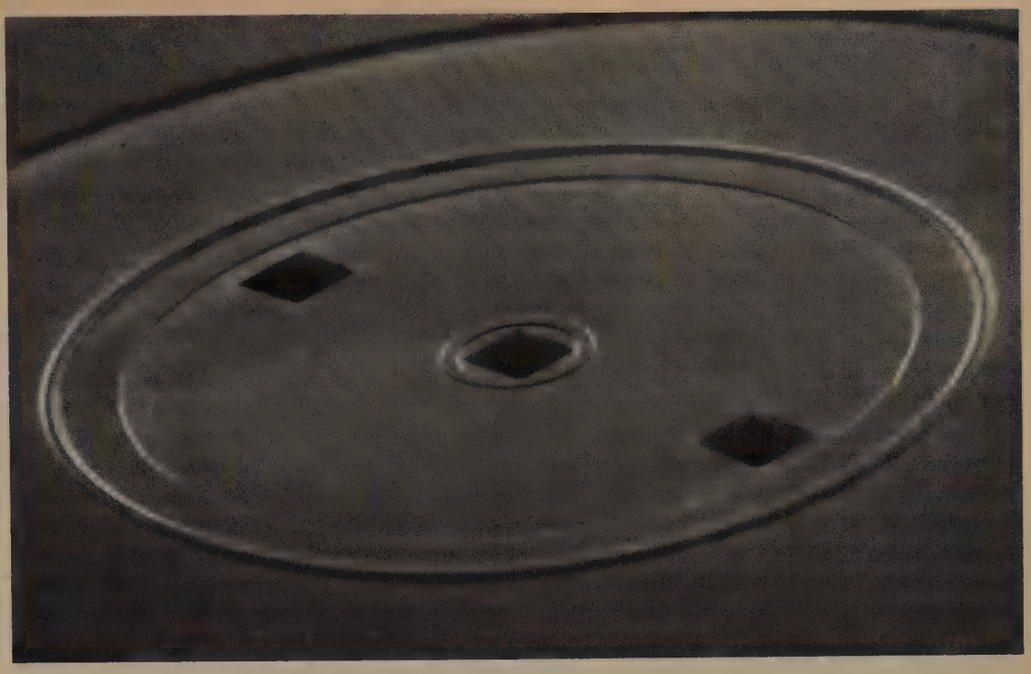
SAMPLE INJECTION VALVE for the miniature chromatograph is fabricated on the same wafer as the capillary column. Two isotropic etching steps define the valve sealing ring and the valve seat, and an anisotropic etch defines the feedthrough holes within the valve well. The first isotropic etch creates a knife-edge sealing ring of silicon raised 10 micrometers above the floor of a well. In the second step a smaller concentric well with another ring—the valve seat—is etched 10 micrometers into the floor of the first well. In the third step the feedthrough holes are formed by an anisotropic etch through square openings in the oxide on the bottom surface of the wafer. The hole within the valve seat opens to the capillary column; the other two holes open to the sample-gas channels. A metal-and-Teflon diaphragm and a solenoid plunger complete the valve. Normally the diaphragm presses against the valve seat, preventing the sample gas from flowing into the column. When the solenoid is actuated, the diaphragm relaxes, allowing pressurized sample gas to flow over the valve seat.

the shallow capillary column might be blocked by extruded glue or solder.

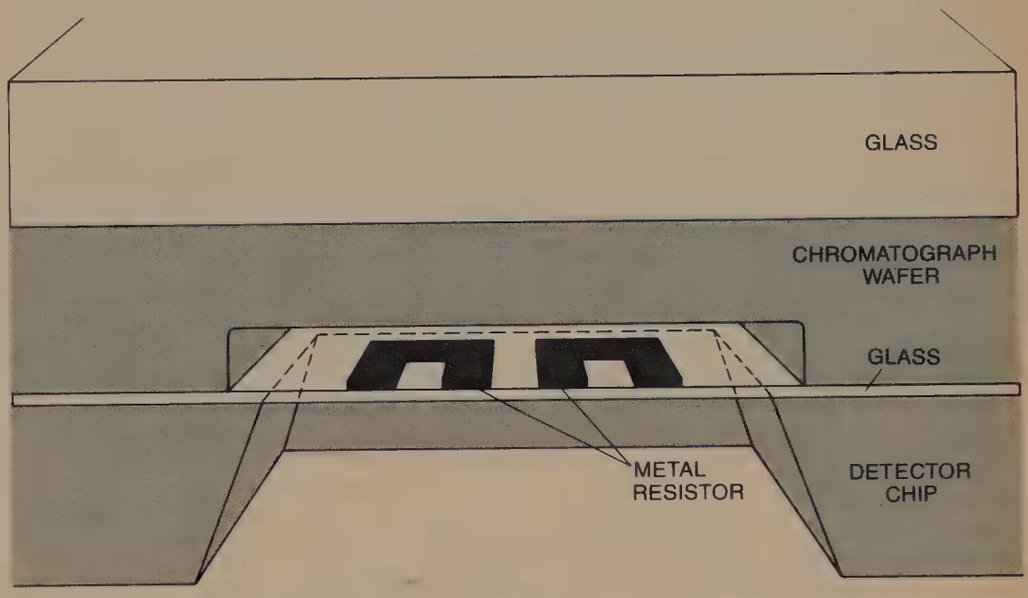
Anodic bonding, which is sometimes called electrostatic bonding or the Mallory process, was developed by George Wallis and Daniel I. Pomerantz of P. R. Mallory & Co. The glass plate and the silicon wafer are placed in contact and heated to about 400 degrees C. A large negative voltage is applied to the top surface of the glass and the silicon is electrically grounded. The glass contains a small amount of sodium; at this temperature the sodium ionizes and becomes mobile. The positive sodium ions move toward the negative electrode, leaving behind bound negative charges in the glass near the gap between the glass and the silicon. As a result the voltage across the gap is nearly equal to the applied voltage. Electrostatic attraction pulls the glass and the silicon into intimate contact and they fuse together, creating a tight seal.

The thermal-conductivity detector developed by John H. Jerman of Stanford for the gas chromatograph is a chip fabricated on a separate wafer. The wafer is oxidized and covered with a layer of Pyrex glass deposited by the technique called sputtering. Thin-film metal resistors are deposited on top of the glass. Then an anisotropic etch through square openings in the oxide on the back of the wafer removes the entire thickness of silicon, leaving a membrane of thermally isolating glass under each set of resistors. The wafer is sawed into chips, to which wires are attached, and the front surface of a chip is clamped over the gas channel at the end of the capillary column.

Because the capillary column has such a small volume, the miniature gas chromatograph requires much smaller amounts of carrier gas than a conventional chromatograph. This and the small size of the unit itself make it possible to build portable chromatography instruments. Indeed, a portable instrument consisting of five miniature chromatographs and a microcomputer is being developed by Microsensor Technology. If the five capillary columns are lined with different materials, the instrument can be made sensitive to about 100 gases. Such an instrument can monitor air quality or measure the energy content of the gas in a natural-gas pipeline.



VALVE WELL shown in this scanning electron micrograph is four millimeters in diameter and surrounds two concentric silicon rings: the valve seal (outer complete ring) and the valve seat (inner complete ring). The ring walls are 50 micrometers thick. The orifice in the center leads to the capillary column and the other two openings lead to the input and exhaust channels.



THERMAL-CONDUCTIVITY DETECTOR for the chromatograph is a resistor defined by etching a thin film of metal deposited on a wafer that has been coated with glass. After the resistor is formed the wafer is etched anisotropically from the back, leaving only a thin membrane of glass under the resistor. Because glass is a poor thermal conductor (whereas silicon is a good one), the membrane thermally isolates the resistor. A change in the thermal conductivity of the gas stream flowing over the resistor alters its temperature and hence its electrical resistance.

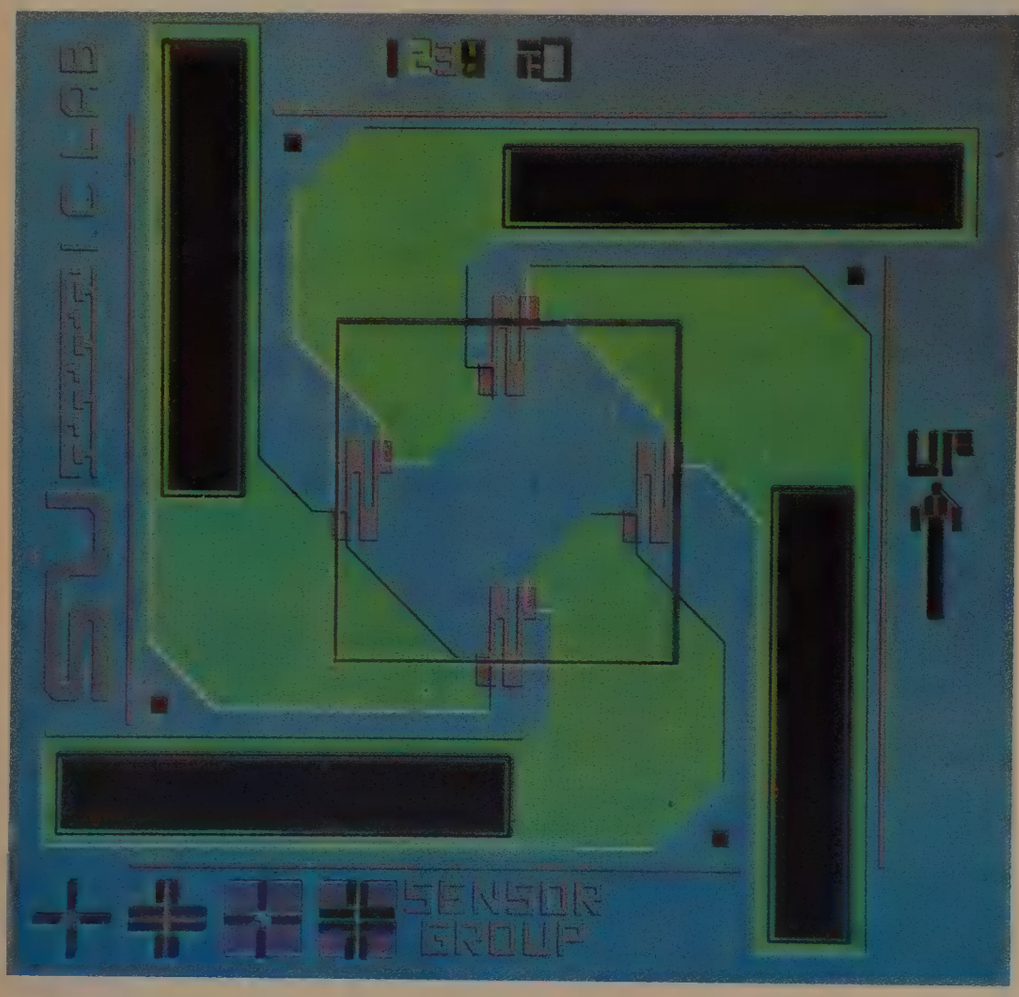
Micromachining has recently been used to fabricate a high-performance heat sink in a silicon wafer. The heat sink is designed to improve the performance of very-high-speed integrated circuits. In general the power dissipation of an electronic device increases along with its speed, and the heat evolved by the fastest circuits would disrupt their own operation if they were not adequately cooled. Even with special packaging schemes for high-speed circuits, the upper limit on power dissipation is about 20 watts per square centimeter of chip.

David B. Tuckerman and Fabian Pease of the Stanford Integrated Circuits Laboratory have developed a heat sink that can handle much higher power densities. The idea is to fabricate the high-speed electronic devices on the front surface of a wafer and to cut closely spaced rectangular grooves in the back of the wafer. A liquid coolant is forced through these channels. Silicon itself has very high thermal conductivity, and Tuckerman and Pease have fabricated devices in which the wall thickness, channel width and channel depth give the highest cooling efficiency.

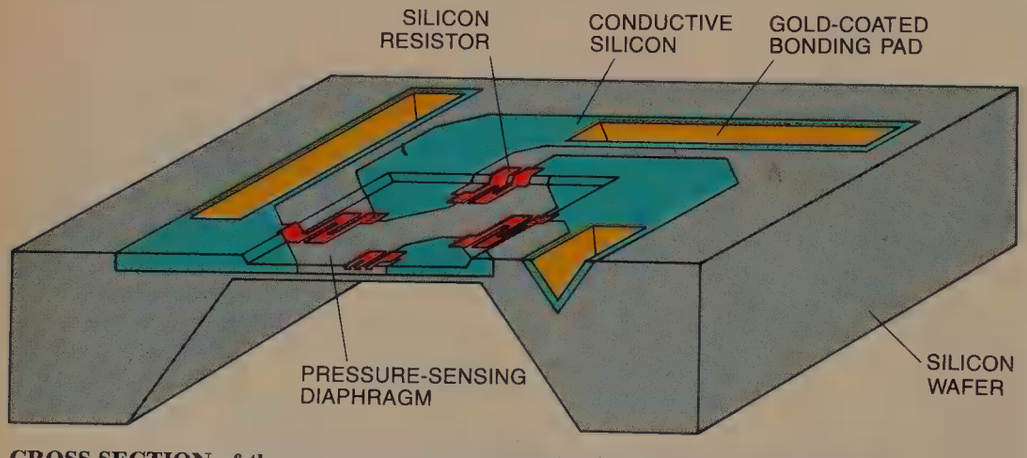
The rectangular grooves are made in much the same way as the *V*-shaped grooves described above. An anisotropic etch through an oxide pattern consisting of narrow parallel lines creates *V* grooves if the wafer, like those discussed so far, is a $<100>$ wafer. If the wafer is a $<110>$ wafer, on the other hand, an anisotropic etch through the same oxide pattern can create grooves with vertical side walls and a roughly rectangular cross section. The rectangular grooves are preferable for a heat sink because they can be spaced more closely than *V* grooves and they expose more surface area to the circulating coolant. After the

grooves are etched the wafer is anodically bonded to glass to create enclosed microchannels. A heat sink with etched channels 300 micrometers deep and 50 micrometers wide, separated by silicon fins 100 micrometers wide, cooled by the forced circulation of water, can handle power dissipations of more than 1,000 watts per square centimeter.

The etched heat sink can be constructed only on silicon and only on wafers of a particular orientation. Tuckerman and Pease have also investigated the possibility of creating heat sinks by sawing grooves rather than etching them. A rotary saw blade with a thin metal rim coated with diamond grit can cut channels from 50 to 200 micrometers wide on the back of a wafer. Although the dimensions of the channels cannot be controlled as precisely with sawing as they can be with etching, sawing does make it possible to create heat sinks on silicon wafers of any orientation and on wafers of other materials, such as the compound semiconductor gallium arsenide and sapphire, a substrate for some types of integrated circuit. In addition the wafers can be sawed in a crosshatch pattern, so that the circuits are supported by micropillars rather than microfins and the heat sink is less likely to be clogged by particles in the coolant.



SILICON PRESSURE SENSOR is shown in the middle of its fabrication sequence. The recessed square in the center of the photomicrograph is a thin pressure-sensing diaphragm .5 millimeter on a side. Purple serpentine elements on the diaphragm are resistors made by diffusing boron into the wafer. Green regions, which act as conductors between the resistors, have a higher boron concentration. Black rectangles near the edge of the sensor are incomplete bonding pads. They are deep *V*-shaped grooves that will be coated with gold, to which lead wires can be attached. When the diaphragm flexes, the resistance of the purple areas changes as a result of the mechanical stress, an effect called piezoresistance. The resistors are positioned so that the resistance of two of them increases and that of the other two decreases; the two-legged resistor configuration is twice as sensitive to pressure changes as a single resistor would be.



CROSS SECTION of the pressure sensor exaggerates the thickness of the diaphragm in order to show detail. The diaphragm is defined by a shallow anisotropic etch from the top of the wafer and a deep etch from the bottom. When the sensor is complete, the top surface is bonded to a glass plate, creating a sealed cavity. The bottom of the diaphragm is exposed to outside pressure. Since the pressure in the cavity is known, the output is a measure of absolute pressure.

Accelerometers and pressure sensors are two examples of integrated silicon sensors. The smallest accelerometer fabricated up to now is an oxide-cantilever device with an on-chip amplifier. The accelerometer, designed by Petersen, Anne C. Shartel and Norman F. Raley of the IBM San Jose Research Laboratory, consists of an oxide beam suspended over a shallow well formed by the boron etch-stop technique. A metal layer is deposited on the top surface of the oxide cantilever. The metal layer and the flat silicon on the bottom of the well act as two plates of a variable air-gap capacitor. A lump of gold is formed on the free end of the beam by plating. If the silicon chip is moved suddenly, the inertia of the gold weight causes the beam to flex, changing the air gap and hence the capacitance.

The output of the sensor is a voltage proportional to acceleration. The instrument has a sensitivity of two millivolts per *g* (the unit of acceleration equal to the acceleration of gravity at the earth's surface). It should be noted that the amplifier is an indispensable part of the device. The change in capacitance being detected is very small and would be swamped by larger capacitance changes in any cables attached to the sensor if the signal were not amplified on the chip. This is true in general of small capacitive sensors; signal conditioning of some kind must precede transmission.

The most successful silicon sensors at present are pressure sensors, some of which include integrated electronic devices. There are two basic types of silicon pressure sensors: piezoresistive and capacitive sensors. The electrical resistors on the thin flexible diaphragm of a piezoresistive sensor change resistance when the diaphragm flexes. In a capacitive sensor the flexible diaphragm is one plate of a variable air-gap capacitor.

In most cases the diaphragm of a pressure sensor is anisotropically etched in a $<100>$ silicon wafer. A simple timed etch can be used to make diaphragms from 10 to 50 micrometers thick. Thin-

ner diaphragms must be made from etch-stop layers.

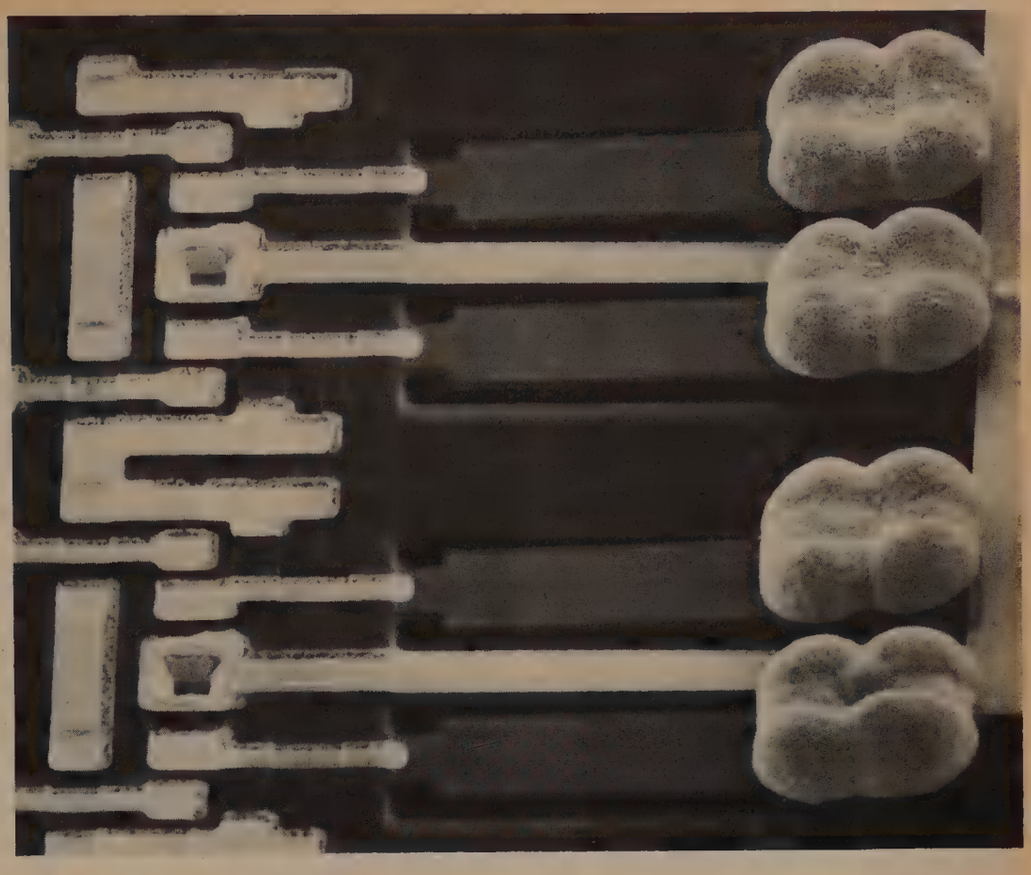
The resistors in a piezoresistive sensor are formed by doping some areas of the diaphragm. Two types of dopants are available. Elements to the right of silicon in the periodic table, such as phosphorus, have five valence electrons, whereas silicon has four. The extra electron can become a carrier of electric current. Elements to the left of silicon in the periodic table, such as boron, have three valence electrons. The absence of an electron (a "hole") can also act as a current carrier, although holes are less mobile than electrons. Current will not flow from a phosphorus-doped region to a boron-doped region because of the difference in charge carriers. Thus if the silicon from which the wafer was made was lightly doped with phosphorus, a heavier concentration of boron can be diffused into selected areas of the wafer to make resistors that are electrically isolated from the rest of the diaphragm.

The piezoresistive effect is a change in electrical resistance resulting from mechanical stress; the magnitude of the effect depends on the orientation of the resistors. By fortunate coincidence boron-doped resistors in a $<100>$ wafer are most sensitive to stress when they are aligned with the $<110>$ edges of an anisotropically etched cavity.

A piezoresistive pressure sensor we developed has four resistors on a square diaphragm. The resistors are near the diaphragm edges (the areas of maximum stress) and are aligned with the $<110>$ directions (the directions along which response to stress is greatest). The resistors are joined in the two-legged configuration known as a Wheatstone bridge. When the diaphragm flexes, the resistance of two resistors increases and that of the other two decreases by an equal amount. As a result the bridge is twice as sensitive to pressure changes as a single resistor would be.

The first pressure-sensor chip to include active circuitry (that is, circuitry including transistors that amplify the signal) was designed by John M. Borky and Kensall Wise of the University of Michigan. The circuitry converts the output voltage of a Wheatstone bridge into a frequency-modulated (FM) signal; in other words, changes in pressure are translated into changes in frequency. If a different range of frequencies is assigned to each of several sensors, they can share a transmission line in the same way different FM radio stations share the radio-frequency band.

Capacitive pressure sensors are inherently more sensitive to pressure changes than piezoresistive sensors, but as we pointed out above, the signal from a capacitive sensor can easily be lost in noise. Signal-conditioning circuitry improves the performance of a piezoresistive sensor, but it is crucial to that



CAPACITIVE ACCELEROMETERS are cantilever beams of silicon dioxide overhanging shallow wells in the silicon. Four accelerometers are shown in this scanning electron micrograph. Bumps of gold at the right end of each cantilever increase the inertia of the beam and therefore the sensitivity of the device. A metal layer on top of the beam, an electrically conductive boron etch-stop layer at the bottom of the well and the air gap between them form a variable capacitor. Under acceleration the beam flexes, changing the capacitance. The circuitry at the left amplifies the signal; the sensor output is a voltage proportional to acceleration.

of a capacitive sensor. Craig S. Sander, James W. Knutti and James D. Meindl of the Stanford Integrated Circuits Laboratory designed the first capacitive pressure sensor with on-chip active circuitry. The diaphragm of this sensor, which is recessed about five micrometers into the silicon wafer, serves as one plate of a capacitor. The other plate is a metal film on a sheet of glass that is anodically bonded to the wafer.

A fixed quantity of gas is sealed in the cavity between the diaphragm and the metal-coated glass plate. The other side of the diaphragm is exposed to the pressure to be measured. When the pressure changes and the diaphragm flexes, the thickness of the insulating air gap between the metal and the silicon changes. The active circuitry on the chip monitors the resulting changes in capacitance, putting out a series of voltage pulses. The time between two pulses is proportional to the measured pressure. This pulse-time-coded signal, like the FM signal from the integrated piezoresistive pressure sensor, is less easily obscured by noise than a simple uncoded voltage would be.

Integrated sensors can meet the requirements of a wider variety of applications than discrete sensors because the signals are relatively immune to noise and leakage. Thus integrated sensors can be put in harsh environments, such as under the hood of an automobile, or in situations where the sensing device must be physically separated from the signal-processing electronics, as in biomedical monitoring. Many of the sensors now under development are designed for such difficult applications. For example, two of us (Barth and Angell) are working on sensors to be implanted in living tissue.

Also under development are chips with several types of sensors, such as a temperature and a pressure sensor, several copies of the same sensor (for reliability) and chips that include some means of sensor recalibration. In the future many integrated sensors will be able to convert analogue signals into digital signals so that they can be connected directly to digital electronics. In general the biggest challenge at present is to determine how much circuitry it is feasible and cost-effective to put on a sensor chip.

At the same time that a great deal of attention is being given to sensors, strictly mechanical silicon devices continue to be developed. Silicon etching techniques are still being refined, and new micromechanical structures will emerge from the laboratories over the next several years. Silicon mechanical devices may one day be nearly as commonplace as silicon electronic devices are now.

The Structure of Quarks and Leptons

They have been considered the elementary particles of matter, but instead they may consist of still smaller entities confined within a volume less than a thousandth the size of a proton

by Haim Harari

In the past 100 years the search for the ultimate constituents of matter has penetrated four layers of structure. All matter has been shown to consist of atoms. The atom itself has been found to have a dense nucleus surrounded by a cloud of electrons. The nucleus in turn has been broken down into its component protons and neutrons. More recently it has become apparent that the proton and the neutron are also composite particles; they are made up of the smaller entities called quarks. What comes next? It is entirely possible that the progression of orbs within orbs has at last reached an end and that quarks cannot be more finely divided. The leptons, the class of particles that includes the electron, could also be elementary and indivisible. Some physicists, however, are not at all sure the innermost kernel of matter has been exposed. They have begun to wonder whether the quarks and leptons too might not have some internal composition.

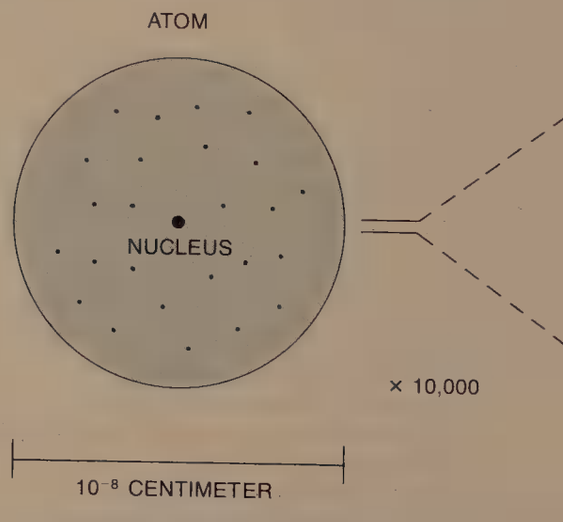
The main impetus for considering still another layer of structure is the conviction (or perhaps prejudice) that there should be only a few fundamental building blocks of matter. Economy of means has long been a guiding principle of physics, and it has served well up to now. The list of the basic constituents of matter first grew implausibly long toward the end of the 19th century, when the number of chemical elements, and hence the number of species of atoms, was approaching 100. The resolution of atomic structure solved the problem, and in about 1935 the number of elementary particles stood at four: the proton, the neutron, the electron and the neutrino. This parsimonious view of the world was spoiled in the 1950's and 1960's; it turned out that the proton and the neutron are representatives of a very large family of particles, the family now called hadrons. By the mid-1960's the number of fundamental forms of matter was again roughly 100. This time it was the quark model that brought relief. In the initial formulation of the model all hadrons could be explained as combinations of just three kinds of quarks.

Now it is the quarks and leptons themselves whose proliferation is beginning to stir interest in the possibility of a simpler scheme. Whereas the original model had three quarks, there are now thought to be at least 18, as well as six leptons and a dozen other particles that act as carriers of forces. Three dozen basic units of matter are too many for the taste of some physicists, and there is no assurance that more quarks and leptons will not be discovered. Postulating a still deeper level of organization is perhaps the most straightforward way to reduce the roster. All the quarks and leptons would then be composite objects, just as atoms and hadrons are, and would owe their variety to the number of ways a few smaller constituents can be brought together. The currently observed diversity of nature would be not intrinsic but combinatorial.

It should be emphasized that as yet there is no evidence quarks and leptons have an internal structure of any kind. In the case of the leptons, experiments have probed to within 10^{-16} centimeter and found nothing to contradict the assumption that leptons are pointlike and structureless. As for the quarks, it has not been possible to examine a quark in isolation, much less to discern any possible internal features. Even as a strictly theoretical conception, the subparticle idea has run into difficulty: no one has been able to devise a consistent description of how the subparticles might move inside a quark or a lepton and how they might interact with one another. They would have to be almost unimaginably small: if an atom were magnified to the size of the earth, its innermost constituents could be no larger than a grapefruit. Nevertheless, models of quark and lepton substructure make a powerful appeal to the aesthetic sense and to the imagination: they suggest a way of building a complex world out of a few simple parts.

Any theory of the elementary particles of matter must also take into account the forces that act between them and the laws of nature that govern the forces. Little would be gained in simplifying the spectrum of particles if the number of forces and laws were thereby increased. As it happens, there has been a subtle interplay between the list of particles and the list of forces throughout the history of physics.

In about 1800 four forces were thought to be fundamental: gravitation, electricity, magnetism and the short-range force between molecules that is responsible for the cohesion of matter. A series of remarkable experimental and theoretical discoveries then led to the recognition that electricity and magnetism are actually two manifestations of the same basic force, which was soon given the name electromagnetism. The discovery of atomic structure brought a further revision. Although an atom is



HIERARCHY OF PARTICLES in the structure of matter currently has four levels. All matter is made up of atoms; the atom con-

electrically neutral overall, its constituents are charged, and the short-range molecular force came to be understood as a complicated residual effect of electromagnetic interactions of positive nuclei and negative electrons. When two neutral atoms are far apart, there are practically no electromagnetic forces between them. When they are near each other, however, the charged constituents of one atom are able to "see" and influence the inner charges of the other, leading to various short-range attractions and repulsions.

As a result of these developments physics was left with only two basic forces. The unification of electricity and magnetism had reduced the number by one, and the molecular interaction had been demoted from the rank of a fundamental force to that of a derivative one. The two remaining fundamental forces, gravitation and electromagnetism, were both long-range. The exploration of nuclear structure, however, soon introduced two new short-range forces. The strong force binds protons and neutrons together in the nucleus, and the weak force mediates certain transformations of one particle into another, as in the beta decay of a radioactive nucleus. Thus there were again four forces.

The development of the quark model and the accompanying theory of quark interactions was the next occasion for revising the list of forces. The quarks in a proton or a neutron are thought to be held together by a new long-range fundamental force called the color force, which acts on the quarks because they bear a new kind of charge called color. (Neither the force nor the charge has any relation to ordinary colors.) Just as an atom is made up of electrically charged constituents but is itself neutral, so a proton or a neutron is made up of colored quarks but is itself colorless. When two colorless protons are far apart, there are essentially no color forces between them, but when they are near, the colored quarks in one proton "see" the color charges in the other proton. The short-range attractions and repulsions that result have been identified with the effects of the strong force. In other words, just as the short-range molecular force became a residue of the long-range electromagnetic force, so the short-range strong force has become a residue of the long-range color force.

One more chapter can be added to this abbreviated history of the forces of nature. A deep and beautiful connection has been found between electromagnetism and the weak force, bringing them almost to the point of full unification. They are clearly related, but the connection is not quite as close as it is in the case of electricity and magnetism, and so they must still be counted as separate forces. Therefore the current list of fundamental forces still has four entries: the long-range gravitational, electromagnetic and color forces and the short-range weak force. Within the limits of present knowledge all natural phenomena can be understood through these forces and their residual effects.

The evolution of ideas about particles and that of ideas about forces are clearly interdependent. As new basic particles are found, old ones turn out to be composite objects. As new forces are discovered, old ones are unified or reduced to residual status. The lists of particles and forces are revised from time to time as matter is explored at smaller scale and as theoretical understanding progresses. Any change in one list inevitably leads to a modification of the other. The recent speculations about quark and lepton structure are no exception; they too call for changes in the complement of forces. Whether the changes represent a simplification remains to be seen.

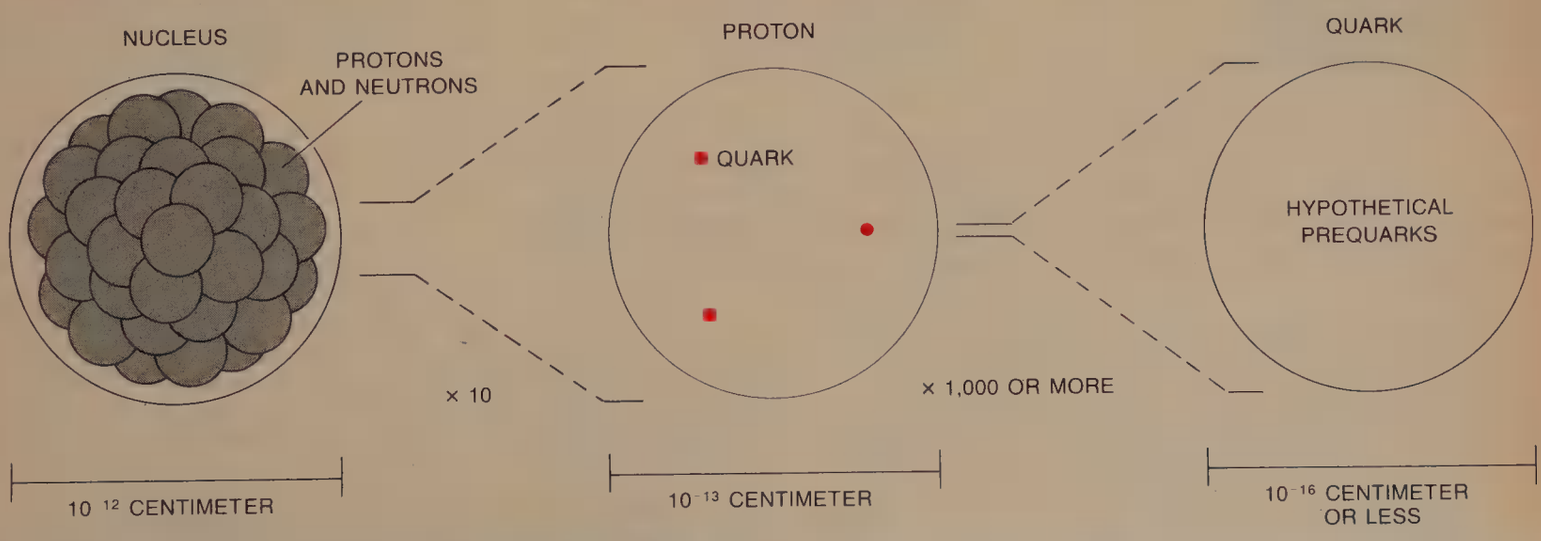
Of the four established fundamental forces, gravitation must be put in a category apart. It is too feeble even to be detected in the interactions of individual particles, and it is not understood in terms of microscopic events. For the other three forces successful theories have been developed and are now widely accepted. The three theories are distinct, but they are consistent with one another; taken together they constitute a comprehensive model of elementary particles and their interactions, which I shall refer to as the standard model.

In the standard model the indivisible constituents of matter are the quarks and the leptons. It is convenient to discuss the leptons first. There are six of them: the electron and its companion the electron-type neutrino, the muon and the muon-type neutrino and the tau and the tau-type neutrino. The electron, the muon and the tau have an electric charge of -1; the three neutrinos are electrically neutral.

There are also six basic kinds of quark, which have been given the names up, down, charmed, strange, top and bottom, or *u, d, c, s, t* and *b*. (The top quark has not yet been detected experimentally, and neither has the tau-type neutrino, but few theorists doubt their existence.) The *u, c* and *t* quarks have an electric charge of $+2/3$, the *d, s* and *b* quarks a charge of $-1/3$. In addition each quark type has three possible colors, which I shall designate red, yellow and blue. Thus if each colored quark is counted as a separate particle, there are 18 quark varieties altogether. Note that each quark carries both color and electric charge, but none of the leptons are colored.

For each particle in this scheme there is an antiparticle with the same mass but with opposite values of electric charge and color. The antiparticle of the electron is the positron, which has a charge of $+1$. The antiparticle of a red *u* quark, with a charge of $+2/3$, is an antired $\bar{u}$ antiquark, with a charge of $-2/3$.

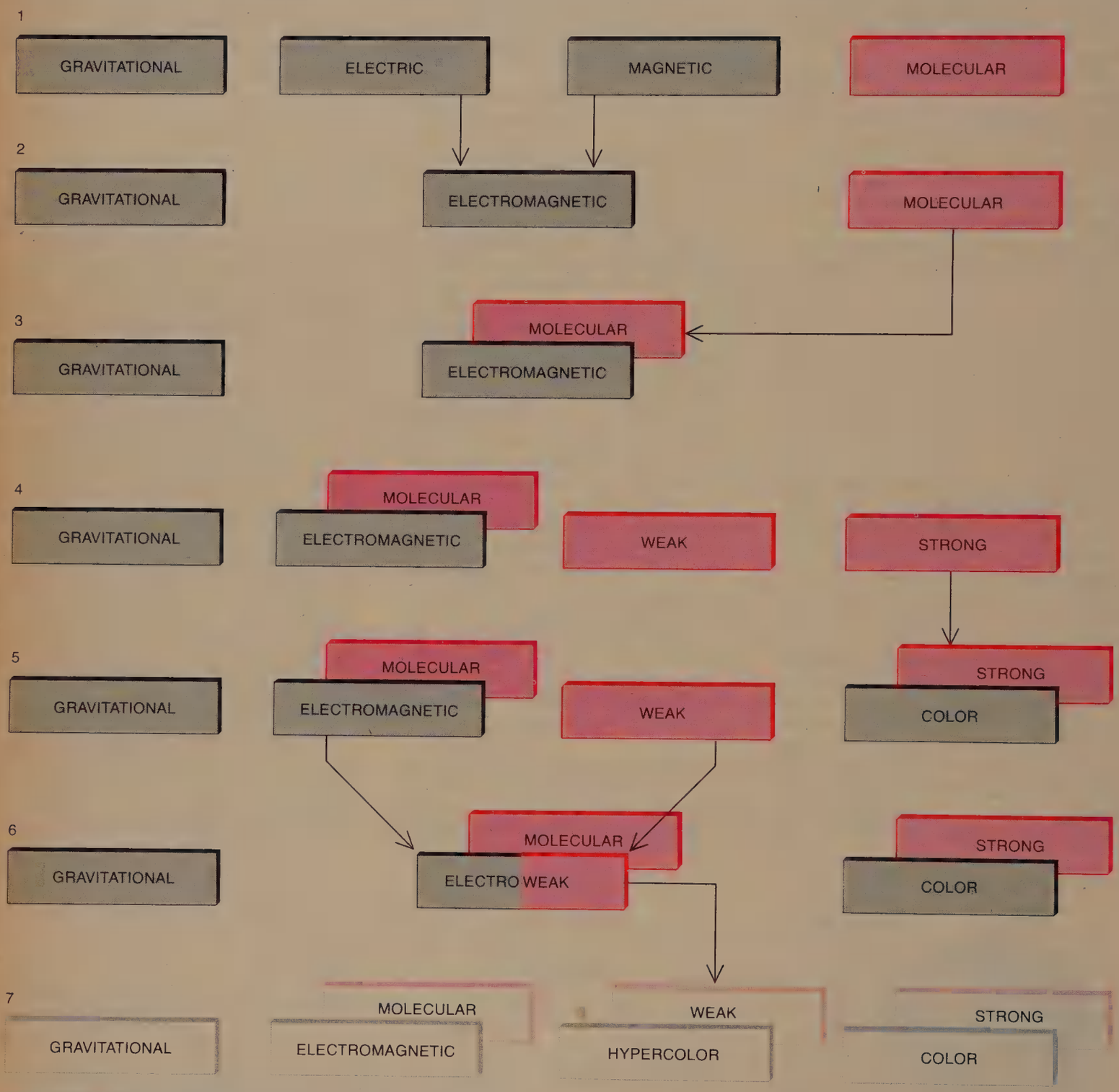
The color property of the quarks is analogous in many ways to electric charge, but because there are three pos-



sists of a nucleus surrounded by electrons; the nucleus is composed of protons and neutrons; each proton and neutron is thought to be composed of three quarks. Recent speculations might add a fifth level: the quark might be a composite of hypothetical finer constituents, which can be generically called prequarks. The leptons, the class of particles that includes the electron, could also consist of prequarks.

sible colors it is appreciably more complicated. Electrically charged particles can be brought together to form an electrically neutral system in only one way: by combining equal quantities of positive and negative charge. A colorless composite particle can be formed out of colored quarks in much the same way, namely by combining a colored quark and an anticolored antiquark. In the case of color, however, there is a second way to form a neutral state: any composite system with equal quantities of all three colors or of all three anticolors is also colorless. For this reason a proton consisting of one red quark, one yellow quark and one blue quark has no net color.

One further property of the quarks and leptons should be mentioned: each particle has a spin, or intrinsic angular momentum, equal to one-half the basic quantum-mechanical unit of angular momentum. When a particle with a spin of 1/2 moves along a straight line, its intrinsic rotation can be either clockwise or counterclockwise when the particle is viewed along the direction of motion. If the spin is clockwise, the particle is said to be right-handed, because when the fingers of the right hand curl in the same direction as the spin, the thumb indicates the direction of motion. For a par-



FUNDAMENTAL FORCES OF NATURE can be classified in a scheme that has evolved together with the list of elementary particles. Here long-range forces are shown in gray and short-range ones in color. Early in the 19th century three long-range forces were thought to be fundamental: the gravitational, the electric and the magnetic. One short-range force, the molecular force responsible for the cohesion of matter, also had fundamental status. James Clerk Maxwell unified electricity and magnetism, and with the discovery of atomic structure it became apparent that the molecular force is not fundamental but instead is a residual effect of electromagnetism. The discovery of the atomic nucleus introduced two short-range forces, the weak and the strong. In the quark model, however, the strong force becomes a residue of a new long-range force called the color force. Furthermore, a deep relation has been found between the weak force and electromagnetism, so that they can be considered partially unified. The sixth row of the chart represents the forces of nature as they are now understood in the "standard model" of elementary-particle physics. A successful model of quark and lepton structure could bring a further revision. In some models, for example, there is a new long-range force called hypercolor, and the weak force is a residue of it. In models of this kind all the fundamental forces of nature are long-range ones. Such models, however, are still highly speculative.

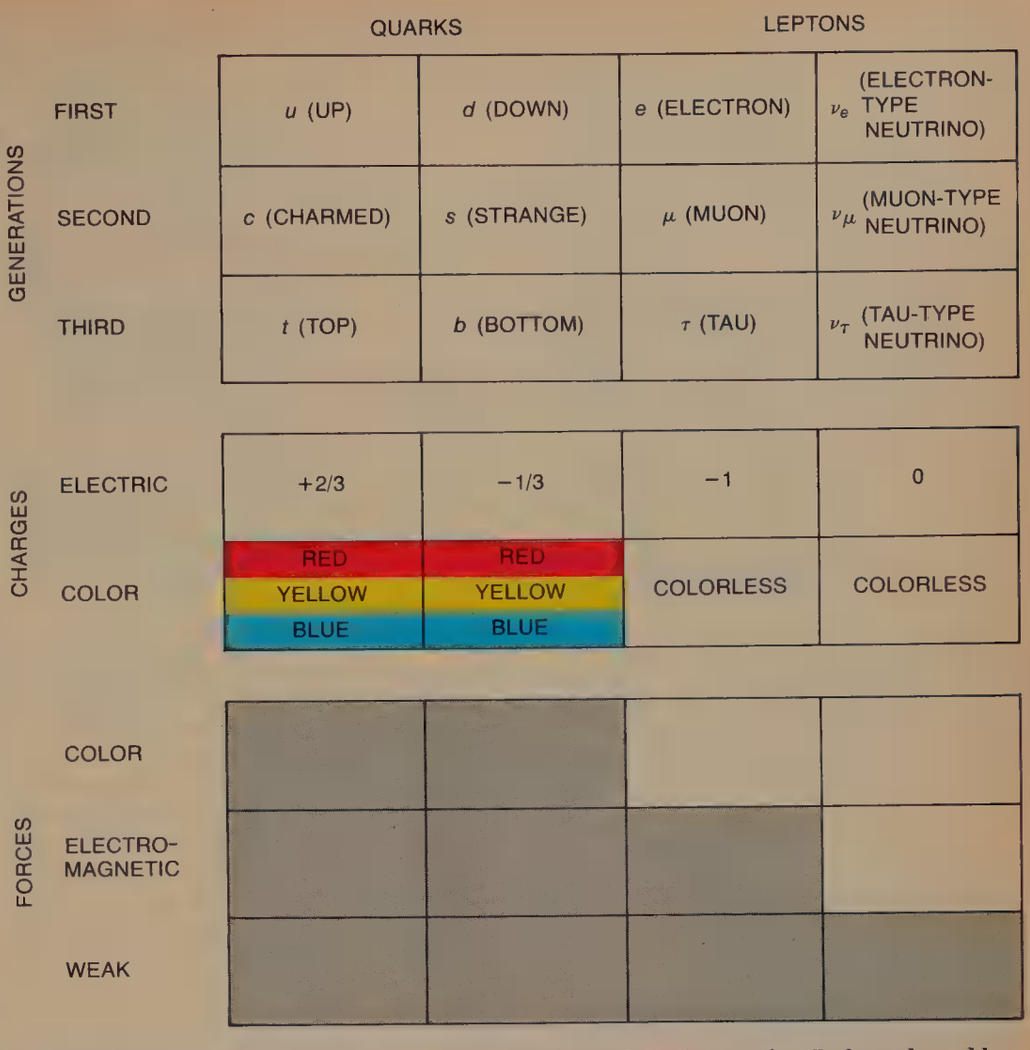
ticle with the opposite sense of spin a left-hand rule describes the motion, and so the particle is said to be left-handed.

In the standard model the three forces that act on the quarks and leptons are described by essentially the same mathematical structure. It is known as a gauge-invariant field theory or simply a gauge theory. Each force is transmitted from one particle to another by carrier fields, which in turn are embodied in carrier particles, or gauge bosons.

The gauge theory of the electromagnetic force, called quantum electrodynamics or QED, is the earliest and simplest of the three theories. It was devised in the 1940's by Richard P. Feynman, Julian S. Schwinger and Sin-Itiro Tomonaga. QED describes the interactions of electrically charged particles, most notably the electron and the positron. There is one kind of gauge boson to mediate the interactions; it is the photon, the familiar quantum of electromagnetic radiation, and it is massless and has no electric charge of its own. QED is probably the most accurately tested theory in physics. For example, it correctly predicts the magnetic moment of the electron to at least 10 significant digits.

The theory of the color force was formulated by analogy to QED and is called quantum chromodynamics or QCD. It was developed over a period of almost two decades through the efforts of many theoretical physicists. In QCD particles interact by virtue of their color rather than their electric charge. The gauge bosons of QCD, which are responsible for binding quarks inside a hadron, are called gluons. Like the photon, the gluons are massless, but whereas there is just one kind of photon, there are eight species of gluons. A further difference between the photon and the gluons turns out to be even more important. Although the photon is the intermediary of the electromagnetic force, it has no electric charge and hence gives rise to no electromagnetic forces of its own (or at least none of significant magnitude). The gluons, in contrast, are not colorless. They transmit the color force between quarks but they also have color of their own and respond to the color force. This reflexiveness, whereby the carrier of the force acts on itself, makes a complete mathematical analysis of the color force exceedingly difficult.

One peculiarity that seems to be inherent in QCD is the phenomenon of color confinement. It is thought that the color force somehow traps colored objects (such as quarks and gluons) inside composite objects that are invariably colorless (such as protons and neutrons). The colored particles can never escape (although they can form new colorless combinations). It is because of color confinement, physicists suppose, that a quark or a gluon has never been seen in isolation. I must stress that although the idea of color confinement is now widely accepted, it has not been proved to follow from QCD. There may still be surprises in store.



STANDARD MODEL of elementary particles includes three "generations" of quarks and leptons, although all ordinary matter can be constructed out of the particles of the first generation alone. The quarks are distinguished by fractional values of electric charge and by a property that is fancifully called color: each quark type comes in red, yellow and blue versions. The leptons have integer units of electric charge and are colorless. The two classes of particles also differ in their response to the various forces. Only the quarks are subject to the color force, and as a result they may be permanently confined inside composite particles such as the proton.

The weak force is somewhat different from the other two, but it can nonetheless be described by a gauge theory of the same general kind. The theory was worked out, and the important connection between the weak force and electromagnetism was established, in the 1960's and the early 1970's by a large number of investigators. Notable contributions were made (in chronological order) by Sheldon Lee Glashow of Harvard University, Steven Weinberg of the University of Texas at Austin, Abdus Salam of the International Centre for Theoretical Physics in Trieste and Gerard 't Hooft of the University of Utrecht.

Curiously, the charges on which the weak force acts are associated with the handedness of a particle. Among both quarks and leptons left-handed particles and right-handed antiparticles have a weak charge, but right-handed particles and left-handed antiparticles are neutral with respect to the weak force. What is odder still, the weak charge is not conserved in nature: a unit of charge can be created out of nothing or can disappear into the vacuum. In contrast, the net quantity of electric charge in an isolated system of particles can never be altered, and neither can the net color. The weak force is also distinguished by its exceedingly short range; its effects extend only to a distance of about 10^{-16} centimeter, or roughly a thousandth of the diameter of a proton.

In the gauge theory of the weak force both the failure of the weak charge to be conserved and the short range of the force are attributed to a mechanism called spontaneous symmetry breaking, which I shall discuss in greater detail below. For now it is sufficient to note that the symmetry-breaking mechanism implies that the weak charge, and the associated handedness of particles, should be conserved at extremely high energy, where a particle's mass is a negligible fraction of its kinetic energy.

Spontaneous symmetry breaking also requires that the gauge bosons of the

weak force be massive particles; indeed, they have masses approximately 100 times the mass of the proton. In the standard model there are three such bosons: two of them, designated W^+ and W^-, carry electric charge as well as weak charge; the third, designated Z^0, is electrically neutral. The large mass of the weak bosons accounts for the short range of the force. According to the uncertainty principle of quantum mechanics, the range of a force is inversely proportional to the mass of the particle that transmits it. Thus electromagnetism and the color force, being carried by massless gauge bosons, are effectively infinite in range, whereas the weak force has an exceedingly small sphere of influence. Spontaneous symmetry breaking has still another consequence: it predicts the existence of at least one additional massive particle, separate from the weak bosons. It is called the Higgs particle after Peter Higgs of the University of Edinburgh, who made an important contribution to the theory of spontaneous symmetry breaking.

In the past 10 years the successes of the standard model have given physicists a good deal of self-confidence. All known forms of matter can be constructed out of the 18 colored quarks and the six leptons of the model. All observed interactions of matter can be explained as exchanges of the 12 gauge bosons included in the model: the photon, the eight gluons and the three weak bosons. The model seems to be internally consistent; no one part is in conflict with any other part, and all measurable quantities are predicted to have a plausible, finite value. Internal consistency is not a trivial achievement in a conceptual system of such wide scope. So far the model is also consistent with all experimental results, that is to say, no clear prediction of the model has yet been contradicted by experiment. To be sure, there are some important predictions that have not yet been fully verified; most notably, the tau-type neutrino, the top quark, the weak bosons and the Higgs particle must be found. The first direct evidence of W bosons was recently reported by a group of experimenters at CERN, the European Laboratory for Particle Physics in Geneva. In the next several years new particle accelerators and more sensitive detecting apparatus will test the remaining predictions of the model. Most physicists are quite certain they will be confirmed.

If the standard model has proved so successful, why would anyone consider more elaborate theories? The primary motivation is not a suspicion that the standard model is wrong but rather a feeling that it is less than fully satisfying. Even if the model gives correct answers for all the questions it addresses, many questions are left unanswered and many regularities in nature remain coincidental or arbitrary. In short, the model itself stands in need of explanation.

The strongest hint of some organizing principle beyond the standard model is the proliferation of elementary particles. The known properties of matter are not so numerous or diverse that 24 particles are needed to represent them all. Indeed, there seems to be a great deal of repetition in the spectrum of quarks and leptons. There are three leptons with an electric charge of -1, three neutral leptons, three quarks with a charge of $+2/3$ and three quarks with a charge of $-1/3$. Everything is triplicated, and for no apparent reason. A world constructed by choosing one particle from each of the four groups would seem to have all the necessary variety.

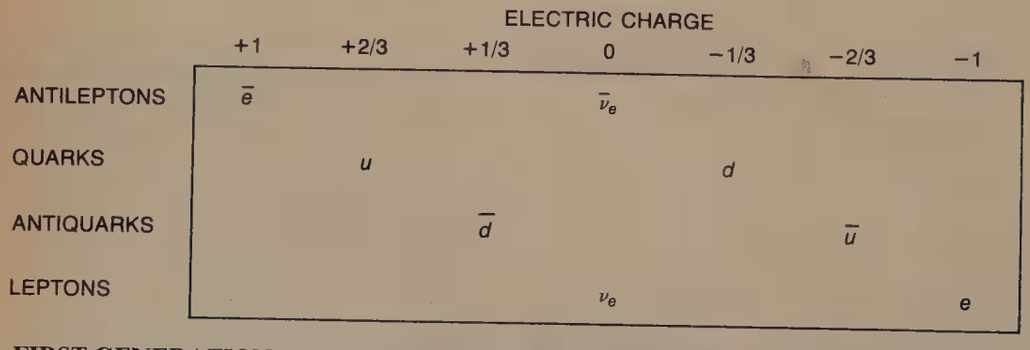
As it turns out, all ordinary matter can indeed be formed from a subset that includes just the u quark, the d quark, the electron and the electron-type neutrino. These four particles and their antiparticles make up the "first generation" of quarks and leptons. The remaining quarks and leptons merely repeat the same pattern in two additional generations without seeming to add anything new. Corresponding particles in different generations are identical in all respects except one: they have different masses. The d, s and b quarks, for example, respond in precisely the same way to the electromagnetic, color and weak forces. For some unknown reason, however, the s quark is roughly 20 times as heavy as the d quark, and the b quark is approximately 600 times as heavy as the d. The mass ratios of the other quarks and of the charged leptons are likewise large and unexplained. (The masses of the neutrinos are too small to have been measured; it is not yet known whether the neutrinos are merely very light or are entirely massless.)

The presence of three generations of quarks and leptons begs for an explanation. Why does nature repeat itself? The pattern of particle masses is also mysterious. In the standard model the masses are determined by approximately 20 "free" parameters that can be assigned any values the theorist chooses; in practice the values are generally based on experimental findings. Is it possible the 20 parameters are all unrelated? Are they fundamental constants of nature with the same status as the velocity of light or the electric charge of the electron? Probably not.

A further tantalizing regularity can be perceived in the electric charges of the quarks and leptons: they are all related by simple ratios and are all integer multiples of one-third the electron charge. The standard model supplies no reason; in principle the charge ratios could have any values. It can be deduced from observation that the ratios of one-third and two-thirds that define the quark charges are not approximations. The proton consists of two u quarks and a d quark, with charges of $2/3 + 2/3 - 1/3$, or $+1$. If these values were not exact and the quarks instead had charges of, say, $+.617$ and $-.383$, the magnitude of the proton's charge would not be exactly equal to that of the electron's, and ordinary atoms would not be electrically neutral. Since atoms can be brought together in enormous numbers, even a slight departure from neutrality could be readily detected.

If the particles and antiparticles that make up a single generation are arranged according to their charge, it is found that every value from -1 to $+1$ in intervals of one-third is occupied by one particle (or, in the case of zero charge, by two particles, namely the neutrino and the antineutrino). The pattern formed raises still more questions. Why has nature favored these values of electric charge but no others, such as $+4/3$ or $-5/3$? It is apparent that all particles with integral charge are colorless and all those with fractional charge are colored. Is there some relation between the electric charge of a particle and its color or between the quarks and the leptons? The standard model implies no such relations, but they seem to exist.

Another motivation for looking beyond the standard model is the continuing desire to unify the fundamental forces, or at least to find some relation



FIRST GENERATION of quarks and leptons forms an orderly pattern when the particles are arranged according to their electric charge. All values of charge from +1 to −1 in intervals of 1/3 are represented. All colored particles have fractional charge and all colorless ones have integral charge. The pattern is an arbitrary feature of the standard model, where charge and color are independent, but it might have some explanation if quarks and leptons are composite.

among them. The cause of parsimony would be served, for example, if two of the forces could be consolidated, as electricity and magnetism were, or if one force could be made a residue of another, as the strong force was made a residue of the color force. Ironically, it may turn out that a simplification of this kind can be attained only by introducing still more forces.

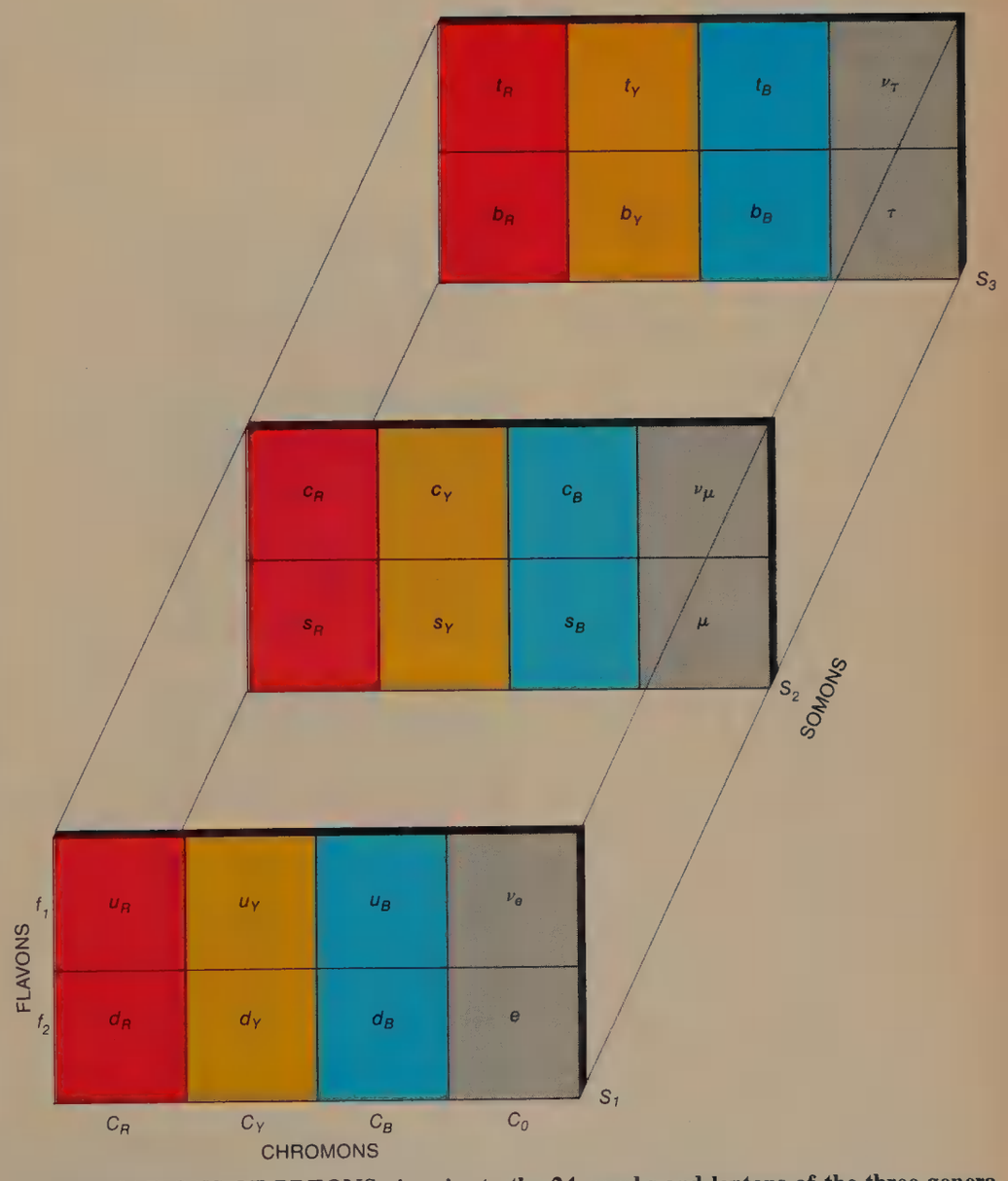
A theory that "goes beyond" the standard model need not contradict or invalidate it. The standard model may emerge as a very good approximation of the deeper theory. The standard model gives a remarkably successful description of all phenomena at distances no smaller than about 10^{-16} centimeter. A deeper theory should therefore focus on events at a still smaller scale. If there are new constituents to be discovered, they must exist within such minuscule regions of space. If there are new forces, their action must be effective only at a distance of less than 10^{-16} centimeter, either because the force is inherently short-range (following the example of the weak force) or because it is subject to some form of confinement (as the color force is).

The search for a theory beyond the standard model was launched almost 10 years ago, and by now several directions have been explored. One promising direction has led to the models known as grand unified theories, which incorporate the electromagnetic, color and weak forces into one fundamental force. The essential idea is to put all the quarks and leptons that make up one generation into a single family; new gauge bosons are then postulated to mediate interactions between the colored quarks and the colorless leptons. The theories account for the regularities noted in the distribution of electric charge and explain the exact commensurability of the quark and lepton charges. On the other hand, they do nothing to reduce the number of fundamental constants, they shed no light on the triplication of the generations and they create certain new theoretical difficulties of their own.

There have been several variations on the theme of grand unification. For example, the concept of horizontal symmetry tackles the triplication problem by establishing a symmetry relation among the generations. The mathematically beautiful idea called supersymmetry relates particles whose spin angular momentum is a half-integer (such as the quarks and leptons) to those with integer spin (such as the gauge bosons). The technicolor theory suggests that the Higgs particle of the standard model is a composite object made up of new fundamental entities; they would be bound together by a new force analogous to the color force and called technicolor. Each of these ideas answers some of the ques-

	PREON	ELECTRIC CHARGE	COLOR	GENERATION NUMBER
FLAVONS	f_1	+1/2	COLORLESS	0
	f_2	−1/2	COLORLESS	0
CHROMONS	C_R	+1/6	RED	0
	C_Y	+1/6	YELLOW	0
	C_B	+1/6	BLUE	0
	C_0	−1/2	COLORLESS	0
SOMONS	S_1	0	COLORLESS	1
	S_2	0	COLORLESS	2
	S_3	0	COLORLESS	3

PREON MODEL assigns three properties of quarks and leptons to three groups of hypothetical constituents called flavons, chromons and somons. A quark or a lepton is formed by choosing one preon from each group. The flavons have the primary responsibility for determining electric charge, the chromons determine color and the somons determine generation number. Ideally each kind of preon would carry just one property, but some adjustment is needed to differentiate the fractional electric charges of the quarks from the integral charges of the leptons. In the version of the model shown here the chromons carry electric charge as well as color.



COMBINATIONS OF PREONS give rise to the 24 quarks and leptons of the three generations. For example, the red s quark is made up of somon S_2 (signifying that the composite is a second-generation particle), in combination with the red chromon and the negative flavon.

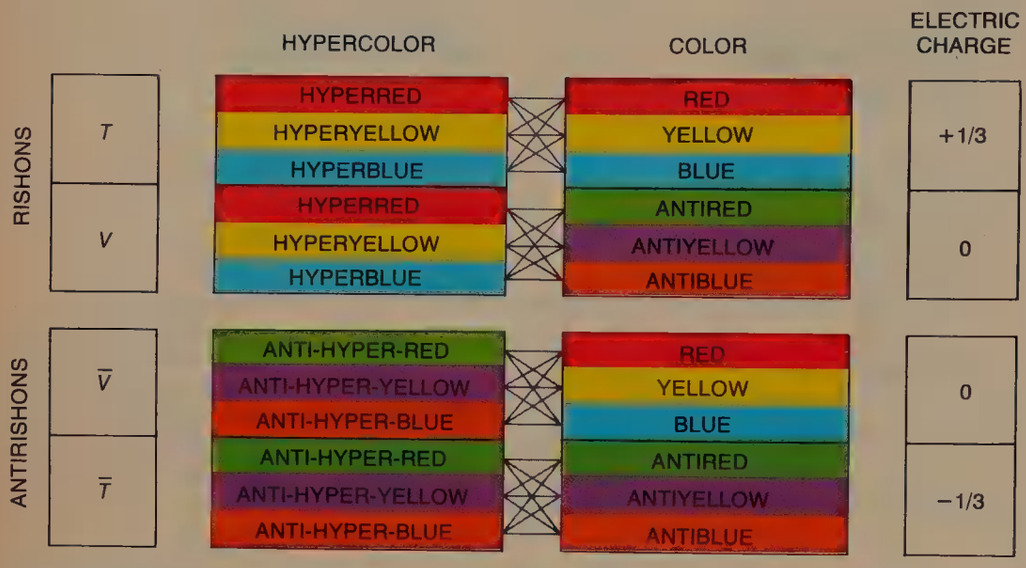
tions that remain open in the standard model. Each idea also fails to answer other questions, raises new difficulties and worsens existing ones, for example by further increasing the number of unrelated arbitrary constants.

In all the above schemes for grand unification it is explicitly assumed that the quarks, the leptons, the photon, the gluons and the weak bosons are the truly fundamental particles of the ultimate theory of nature. The alternative of suggesting that the quarks and leptons are themselves composite is in one sense the most conservative and the least original hypothesis. It is a strategy that has worked before, repeatedly, in going from the atom to the nucleus to the proton to the quark. In another sense the idea of quark and lepton substructure is a most radical proposal. The electron has now been studied for almost a century, and its pointlike nature has been established very well indeed. In the case of the neutrino, which may turn out to be entirely massless, it is even more difficult to imagine an internal structure. The assertion that these particles and the others like them are composites will clearly have to overcome formidable obstacles if it is to have any future.

Offsetting the difficulties of the undertaking are its potential rewards. A fully successful composite model might resolve all the questions left unsettled in the standard model. Such a hypothetical theory would begin by introducing a new set of elementary particles, which I shall refer to generically as prequarks. Ideally there would not be too many of them. Each quark and lepton in the standard model would be accounted for as a combination of prequarks, just as each hadron can be explained as a combination of quarks. The mass of a quark or a lepton would no longer be an arbitrary constant of nature; instead it would be determined by the masses of the constituent prequarks and by the strength of the force that binds the prequarks together. The exact ratios that relate the charge of a quark to that of a lepton would be explained in a similar way: both kinds of composite particles would derive their charges from those of the same constituent prequarks. The entire pattern of quarks and leptons within a generation would presumably reflect some simple rules for combining the prequarks.

The existence of multiple generations might also be explained in a natural way. The quarks and leptons in the higher generations might have an internal constitution similar to that of the corresponding particles of the first generation; the differences could be in the energy and the state of motion of the constituents. Thus the *s* and *b* quarks would be excited states of the *d* quark, and the muon and the tau lepton would be excited states of the electron. Similar excited states are known in all other composite systems, including atoms, nuclei and hadrons. For example, at least a dozen hadrons have been identified in experiments as excited states of the proton; they and the proton itself are all thought to have essentially the same quark composition, namely *uud*.

This imaginary, ideal prequark theory accomplishes everything one might ask of it except for unifying the fundamental forces. Even there some progress is conceivable, since a new force would very likely be introduced to bind the prequarks together; the new force might



RISHON MODEL constructs all the quarks and leptons out of just two species of fundamental particles and their antiparticles. The rishons carry both hypercolor, a property associated with the force that binds them to one another, and ordinary color, which they convey to the composite systems they form. One rishon is electrically charged and the other is neutral.

RISHON COMBINATION	PARTICLE	COLOR	ELECTRIC CHARGE
TTT	$\bar{e}$	COLORLESS	+1
TTV	u	RED, YELLOW, BLUE	+2/3
TVV	$\bar{d}$	ANTIRED, ANTIYELLOW, ANTIBLUE	+1/3
VVV	ν_e	COLORLESS	0
$\bar{V}\bar{V}\bar{V}$	$\bar{\nu}_e$	COLORLESS	0
$\bar{V}\bar{V}\bar{T}$	d	RED, YELLOW, BLUE	−1/3
$\bar{V}\bar{T}\bar{T}$	$\bar{u}$	ANTIRED, ANTIYELLOW, ANTIBLUE	−2/3
$\bar{T}\bar{T}\bar{T}$	e	COLORLESS	−1

COMBINATIONS OF RISHONS taken three at a time give a correct accounting of all the quarks and leptons (and antiquarks and antileptons) in any one generation. The pattern of electric charges noted in the standard model, and the apparent relation between fractional charge and color, emerge as natural consequences of the way the rishons combine. All the allowed combinations of three rishons or of three antirishons are neutral with respect to hypercolor.

lead to a new understanding of how the known forces are related. Imagining what a successful model might be like, however, is not at all the same thing as actually constructing a realistic and internally consistent one. So far no one has done it.

What has been lacking is a satisfactory theory of prequark dynamics, a theory that would describe how the prequarks move inside a quark or a lepton and that would enable one to calculate the mass and total energy of the system. As I shall set forth below, there are fundamental obstacles to the formulation of such a theory, although I would submit that they are not insurmountable. In the meantime, lacking any persuasive account of prequark motions, theorists have nonetheless been exploring the combinatorial possibilities of the prequark idea, that is, they have been examining the ways quarks and leptons might be built up as specific combinations of finer constituents.

In the past few years several dozen composite models have been proposed; they can be classified in perhaps four or five main groups. No single model solves all problems, answers all questions and is widely accepted. It would be unfair to describe only one scheme, but it is impractical to enumerate them all. I shall present a few of the central ideas.

The first explicit model of quark and lepton substructure was proposed in 1974 by Jogesh C. Pati of the University of Maryland at College Park and Salam, who have since returned to the topic several times in collaboration with John Strathdee of the International Centre for Theoretical Physics. It was they who introduced the term prequark, which I have adopted here as a generic name for hypothetical subconstituents of all kinds. The specific elementary particles of the model devised by Pati and Salam I shall call preons, which is another term of their invention.

The rationale for the preon model begins with the observation that every quark and lepton can be identified unambiguously by listing just three of its properties: electric charge, color and generation number. These properties, then, suggest a straightforward way of organizing a set of constituent particles. Three families of preons are needed. In one family the preons carry electric charge, in another they carry color and in the third they have some property that determines generation number. A given quark or lepton is assembled by selecting exactly one preon from each family.

The preons that determine generation number are called somons, from the Greek *soma*, meaning body, because they have a dominant influence on the mass of the composite system. Since there are three generations of quarks and leptons, there must be three somons. The color of the composite system is determined by preons called chromons; there are four of them, one with the color red, one yellow, one blue and one colorless. The remaining family of preons, which is assigned the role of defining electric charge, needs to have only two members in order for every quark and lepton to be uniquely identified. These last preons have been given the name flavons, after flavor, the whimsical term for whatever property it is that distinguishes the u quark from the d quark, the c from the s, the neutrino from the electron and so on.

In the preon model the classification of a composite particle follows directly from its complement of preons. All leptons, for example, are distinguished by a colorless chromon, and all first-generation particles must obviously have a first-generation somon. In the allocation of electric charge, however, a complication arises. If there are only two flavons and if they are the sole carriers of electric charge, not all the charge values observed in nature can be reproduced. The u quark and the neutrino, for example, must have the same charge (because they include the same flavon), and so must the d quark and the electron. The problem can be solved in any of several ways. In one scheme electric charge is assigned to both the flavons and the chromons, and the total charge of a composite particle is equal to the sum of the two values. Models of this kind can be made to yield the correct charge states, but only by abandoning the principle of having each kind of preon carry just one property.

Another troublesome feature of the preon model is the requirement that composites be formed only by drawing one preon from each family. Why are there no particles made up of three chromons, say, or of two somons and a flavon? The exotic properties of such particles would make them quite conspicuous. It seems likely that if they existed, they would have been detected by now.

Many variations of the preon model have been proposed by other physicists, using the same basic idea but slightly different sets of preons. Notable among the variations are the models suggested by Hidezumi Terazawa, Yoichi Chikashige and Keiichi Akama of the University of Tokyo and by O. Wallace Greenberg and Joseph Sucher of the University of Maryland.

Perhaps the simplest model of quark and lepton structure is the rishon model, which I proposed in 1979. A similar idea was put forward at about the same time by Michael A. Shupe of the University of Illinois at Urbana-Champaign. The model has since been further developed and studied in great detail by Nathan Seiberg and me at the Weizmann Institute of Science in Rehovot. The model postulates just two species of fundamental building blocks, called rishons. (*Rishon* is the Hebrew adjective meaning first or primary.) One rishon has an electric charge of $+1/3$ and the other is electrically neutral. I designate them respectively T and V, for *Tohu Vavohu*, Hebrew for "formless and void," the description of the initial state of the universe given in the first chapter of Genesis. The complementary antirishons have charges of $-1/3$ and zero and are designated $\overline{T}$ and $\overline{V}$.

The model has one simple rule for constructing a quark or a lepton: any three rishons can be assembled to form a composite system, or any three antirishons, but rishons and antirishons cannot be mixed in a single particle. The rule gives rise to 16 combinations, which reproduce exactly the properties of the 16 quarks, antiquarks, leptons and antileptons in the first generation. In other words, every quark and lepton in the first generation corresponds to some allowed combination of rishons or antirishons. (In this system of classification each color is counted separately.)

The pattern of quark and lepton charges is generated as follows. The TTT combination, with rishon charges of $1/3 + 1/3 + 1/3$, has a total charge of $+1$ and therefore corresponds to the positron; similarly, $\overline{TTT}$ has a total charge of -1 and is identified with the electron. The VVV and $\overline{VVV}$ combinations are both electrically neutral and represent the neutrino and the antineutrino respectively. The remaining allowed combinations yield fractionally charged quarks. TTV, with a charge of $+2/3$, is the u quark, and TVV, with a charge of $+1/3$, is the $\bar{d}$ antiquark. The analogous antirishon states $\overline{TTV}$ and $\overline{TVV}$ correspond to the d quark and the $\bar{u}$ antiquark.

The model also accounts successfully for the color of the composite systems. A T rishon can have any of the three colors red, yellow and blue, whereas a V rishon has an anticolor. Combinations such as $\overline{TTT}$ and VVV, which designate leptons, can be made colorless since they can include one rishon in each color or one in each anticolor. The other combinations, which yield quarks, must have a net color. For example, a TTV state might have the rishon colors red, blue and antiblue; the antiblue would cancel the blue, leaving the system with a net color of red. In this way the connection between color and electric charge, which was apparent but unexplained in the standard model, is readily understood. Because of the way electric charge and color are allotted to the rishons, all composite systems with fractional charge turn out to be colored, and all systems with an integer charge can be made colorless.

Other regularities of the standard

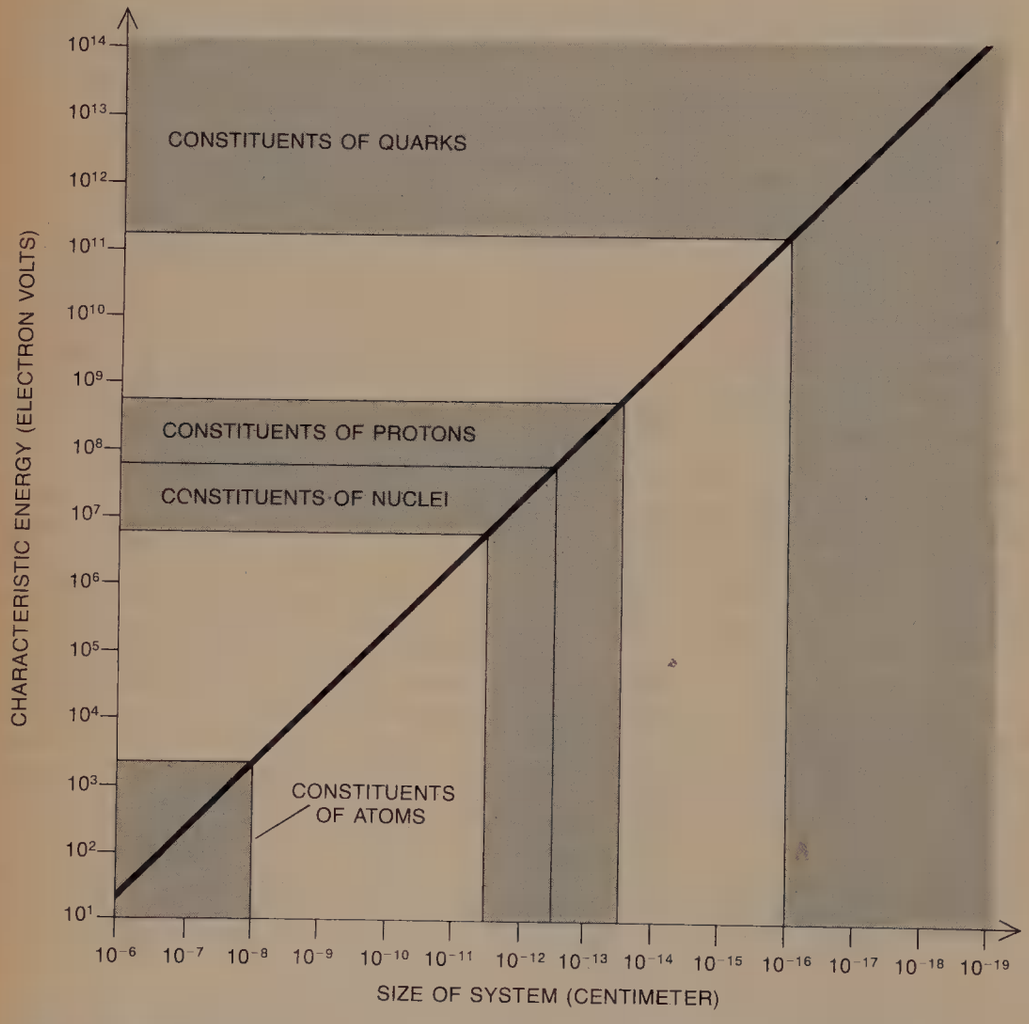
model also lose their air of mystery when rishons are introduced. Consider the hydrogen atom, made up of a proton and an electron, or in terms of quarks and leptons two u quarks, a d quark and an electron. The total rishon content of the quarks is four T's, one $\overline{T}$, two V's and two $\overline{V}$'s. The electric charge of the $\overline{T}$ cancels the charge of one T rishon, and the V's and $\overline{V}$'s also cancel (they have no charge in any case), leaving the proton with a net charge equal to that of a TTT system. The electron's rishon content is just the opposite: $\overline{T}\overline{T}\overline{T}$. Thus it is evident why the proton and the electron have charges of equal magnitude and why the hydrogen atom is neutral: the ultimate sources of the charge are pairs of matched particles and antiparticles.

The rishon model and many other models that explain the pattern of the first generation have difficulty accounting for the second and third generations. It would seem that such models lend themselves well to the scheme of forming each particle in the higher generations as an excited state of the corresponding particle in the first generation. The simplest idea would be to describe the muon, for example, as having the same prequark constituents as the electron, but in the muon the prequarks would have some higher-energy configuration. It is an elegant idea but, regrettably, it appears to be unworkable. The scheme implies differences in energy between the successive excited states that are much larger than the actual differences. The flaw is a fundamental one, and there seems to be no remedy.

Other possible mechanisms for creating multiple generations have been considered. Several physicists have suggested that the higher-generation relatives of a given state might be created by adding a Higgs particle, the "extra" particle associated with the weak bosons in the standard model. Because a Higgs particle has no electric charge or color or even spin angular momentum, adding one to a composite system would alter only the mass. Hence an electron might be converted into a muon by adding one Higgs particle or into a tau by adding two or more Higgs particles. Seiberg and I have proposed another possible mechanism: a higher-generation particle could be formed by the addition of pairs of prequarks and antiprequarks. All charges and other properties must cancel in such a pair, and so again only the mass would be affected.

These ideas are currently at the stage of unrestrained speculation. No one knows what distinguishes the three generations from one another, or why there are three or whether there may be more. No explanation can be given of the mass difference between the generations. In short, the triplication of the generations is still a major unsolved puzzle.

A third kind of substructure model deserves mention. It tries to relate the possibility of quark and lepton structure to another fundamental problem: understanding the relativistic quantum theory of gravitation. Ideas of this kind have been explored by John Ellis, Mary K. Gaillard, Luciano Maiani and Bruno Zumino of CERN. One approach to their ideas is to consider the distances at which prequarks interact: the experimental limit is less than 10^{-16} centimeter, but the actual distance could be several orders of magnitude smaller still. At about 10^{-34} centimeter the gravitational force becomes strong enough to have a significant effect on individual particles. If the scale of the prequark interactions is this small, gravitation cannot be neglected. Ellis, Gaillard, Maiani and Zumino have outlined an ambitious program that aims to unify all the forces, including gravitation, in a scheme that treats not only the quarks and leptons but also the gauge bosons as composite particles. Like other composite models, however, this one has serious flaws.



SIZE AND ENERGY have a reciprocal relation in the quantum theory, indicating that the constituents of composite quarks and leptons must have an exceedingly high kinetic energy. The size of an atom implies that its constituents can have energies ranging from a few electron volts to a few thousand. (An electron volt is the energy gained by an electron accelerated through a potential difference of one volt.) In a nucleus the protons and neutrons move with an energy of several million electron volts, and in a proton or a neutron the quarks have energies of several hundred million electron volts. Any constituents of quarks and leptons must be confined to a radius of less than 10^{-16} centimeter, and possibly much less. As a result the kinetic energy of the hypothetical prequarks can be no less than a few hundred billion electron volts.

Any prequark model, regardless of its details, must supply some mechanism for binding the prequarks together. There must be a powerful attractive force between them. One strategy is to postulate a new fundamental force of nature analogous in its workings to the color force of the standard model. To emphasize the analogy the new force is called the hypercolor force and the carrier fields are called hypergluons. The prequarks are assumed to have hypercolor, but they combine to form hypercolorless composite systems, just as quarks have ordinary color but combine to form colorless protons and neutrons. The hypercolor force presumably also gives rise to the property of confinement, again in analogy to the color force. Hence all hypercolored prequarks would be trapped inside composite particles, which would explain why free prequarks are not seen in experiments. An idea of this kind was first proposed by 't Hooft, who studied some of its mathematical implications but also expressed doubt that nature actually follows such a path.

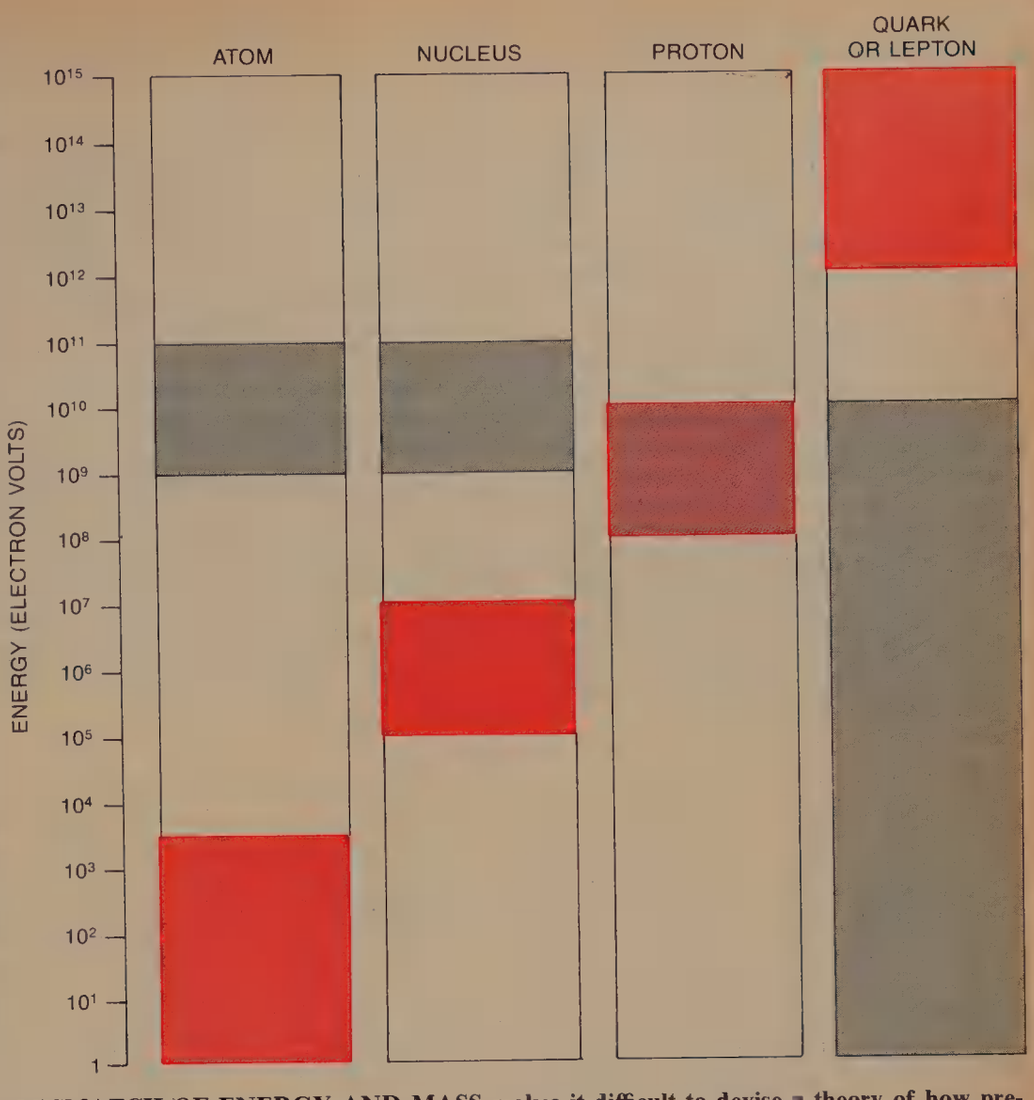
The typical radius of hypercolor con-

finement must be less than 10^{-16} centimeter. Only when matter is probed at distances smaller than this would it be possible to see the hypothetical prequarks and their hypercolors. At a range of 10^{-14} or 10^{-15} centimeter hypercolor effectively disappears; the only objects visible at this scale of resolution (the quarks and leptons) are neutral with respect to hypercolor. At a range of 10^{-13} centimeter ordinary color likewise fades away, and the world seems to be made up entirely of objects that lack both color and hypercolor: protons, neutrons, electrons and so on.

The notion of hypercolor is well suited to a variety of prequark models, including the rishon model. In addition to their electric charge and color the rishons are assumed to have hypercolor and the antirishons to have antihypercolor. Only combinations of three rishons or three antirishons are allowed because only those combinations are neutral with respect to hypercolor. A mixed three-particle system, such as $T\bar{T}$, cannot exist because it would not be hypercolorless. The assignment of hypercolors thereby explains the rule for forming composite rishon systems. Similar rules apply in other hypercolor-based prequark models.

If the aim of a prequark model is to simplify the understanding of nature, postulating a new basic force does not seem very helpful. In the case of hypercolor, however, there may be some compensation. Consider the neutrino: it has neither electric charge nor color, only weak charge. According to the standard model, two neutrinos can act on each other only through the short-range weak force. If neutrinos are composites of hypercolored prequarks, however, there could be an additional source of interactions between neutrinos. When two neutrinos are far apart, there are practically no hypercolor forces between them, but when they are at close range, the hypercolored prequarks inside one neutrino are able to "see" the inner hypercolors of the other one. Complicated short-range attractions and repulsions are the result. The mechanism, of course, is exactly the same as the one that explains the molecular force as a residue of the electromagnetic force and the strong force as a residue of the color force.

The conclusion may also be the same. Seiberg and I, and independently Greenberg and Sucher, were the first to suggest that the short-range weak force may actually be a residual effect of the hypercolor force. According to this hypothesis, the weak bosons W^+, W^- and Z^0 must also be composite objects, presumably made up of certain combinations of the same prequarks that compose the quarks and leptons. If this idea is confirmed, the list of fundamental forces will still have four entries: gravitation, electromagnetism, color and hypercolor. It should be noted, however, that all these forces are long-range ones; the short-range molecular, strong and weak forces will have lost their fundamental status.



MISMATCH OF ENERGY AND MASS makes it difficult to devise a theory of how prequarks might move and interact. In an atom or a nucleus the kinetic energy of the constituents (*color*) is much less than the total mass of the system (*gray*). In a proton the two quantities are of comparable magnitude. In a composite quark, however, the energy of the prequarks greatly exceeds the total mass. Indeed, compared with the kinetic energy, the mass is essentially zero. Somehow virtually all mass is canceled, a development that is unlikely to be accidental.

For now hypercolor remains a conjecture, and so does the notion of explaining the weak force as a residue of the hypercolor force. It may yet turn out that the weak force is fundamental. A careful measurement of the mass, lifetime and other properties of the weak bosons should provide important clues in this matter.

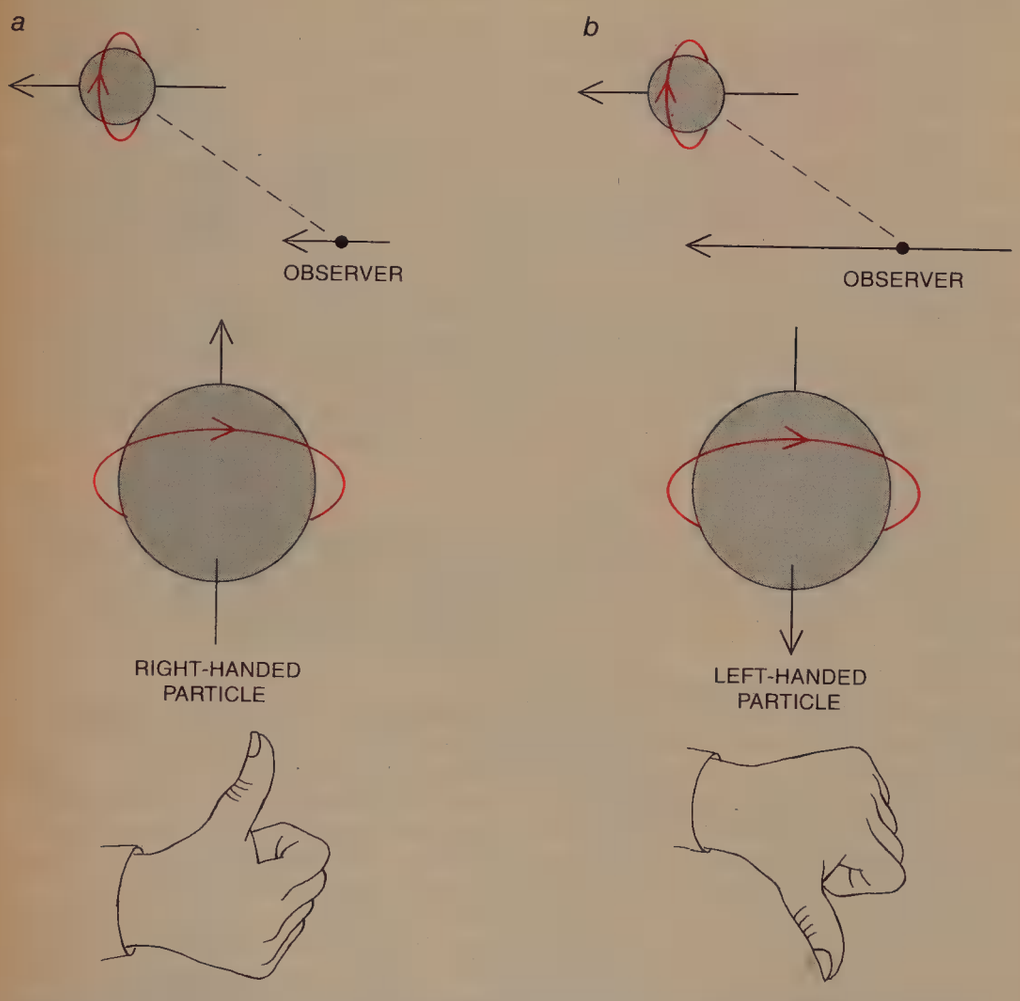
Hypercolor is not the only candidate for a prequark binding force. Another interesting possibility was suggested by Pati, Salam and Strathdee. Instead of introducing a new hypercolor force, they borrowed an idea that has long been familiar, namely the magnetic force, and adapted it to a new purpose. An ordinary magnet invariably has two poles, which can be thought of as opposite magnetic charges. For 50 years there have been theoretical reasons for supposing there could also be isolated magnetic charges, or monopoles. Pati, Salam and Strathdee have argued that the prequarks could be particles with charges resembling both magnetic and electric charges. If they are, the forces binding them may be of a new and interesting origin.

None of the ideas I have just described constitutes a theory of prequark dynamics. Indeed, there is a serious impediment to the formulation of such a theory; it is the requirement that the prequarks be exceedingly small. The most stringent limit on their size is set indirectly by measurements of the magnetic moment of the electron, which agree with the calculations of quantum electrodynamics to an accuracy of 10 significant digits. In the calculations it is assumed that the electron is pointlike; if it had any spatial extension or internal structure, the measured value would differ from the calculated one. Evidently any such discrepancy can at most affect the 11th digit of the result. It is this constraint that implies the characteristic distance scale of the electron's internal structure must be less than 10^{-16} centimeter. Roughly speaking, that is the maximum radius of an electron, and any

prequarks must stay within it. If they strayed any wider, their presence would already have been detected.

Why should the small size of the electron inhibit speculation about its internal structure? The uncertainty principle establishes a reciprocal relation between the size of a composite system and the kinetic energy of any components moving inside it. The smaller the system, the larger the kinetic energy of the constituents. It follows that the prequarks must have enormous energy: more than 100 GeV (100 billion electron volts), and possibly much more. (One electron volt is the energy acquired by an electron when it is accelerated through a potential difference of one volt.) Because mass is fundamentally equivalent to energy, it can be measured in the same system of units. The mass of the electron, for example, is equivalent to .0005 GeV. There is a paradox here, which I call the energy mismatch: the mass of the composite system (if it is indeed composite) is much smaller than the energy of its constituents.

The oddity of the situation can be illuminated by considering the relations of mass and kinetic energy in other composite systems. In an atom the kinetic energy of a typical electron is smaller than the mass of the atom by many orders of magnitude. In hydrogen, for example, the ratio is roughly one part in 100 million. The energy needed to change the orbit of the electron and thereby put the atom into an excited state is likewise a negligible fraction of the atomic mass. In a nucleus the kinetic energy of the protons and neutrons is also small compared with the nuclear mass, but it is not completely negligible. The motion of the particles gives them an energy equivalent to about 1 percent of the system's mass. The energy needed to create an excited state is also about 1 percent of the mass.



CHIRAL SYMMETRY offers a possible explanation of the "miraculous" cancellation of mass in quarks and leptons. Chirality, or handedness, describes the relation of a particle's spin angular momentum to its direction of motion. Suppose an observer is overtaken by a faster-moving electron (*a*). From the observer's point of view the electron obeys a right-hand rule: When the fingers of the right hand curl in the same direction as the spin, the thumb gives the direction of motion. If the observer speeds up, however, so that he overtakes the electron (*b*), the handedness of the particle changes. In the observer's frame of reference the electron is now approaching instead of receding, but its spin direction has not changed; as a result its motion is described by a left-hand rule. Chirality, therefore, is not conserved. There is one kind of particle to which this argument cannot be applied, namely a massless particle, which must always move with the speed of light. No observer can move faster than a massless particle, and so its handedness is an invariant property. If a theory of prequarks had a chiral symmetry, in which handedness must be conserved, the low mass of the quarks and leptons might not be accidental. They would have to be virtually massless for the chiral symmetry to be maintained.

With the proton and its quark constituents the energy-mass relation begins to get curious. From the effective radius of the proton the typical energy of its component quarks can be calculated; it turns out to be comparable to the mass of the proton itself, which is a little less than 1 GeV. The energy that must be invested to create an excited state of the quark system is of the same order of magnitude: the hadrons identified as excited states of the proton exceed it in mass by from 30 to 100 percent. Nevertheless, the ratio of kinetic energy to total mass is still in the range that seems intuitively reasonable. Suppose one knew only the radius of the proton, and hence the typical energy of whatever happens to be inside it, and one were asked to guess the proton's mass. Since the energy of the constituents is generally a few hundred million electron volts, one would surely guess that the total mass of the system is at least of the same order of magnitude and possibly greater. The guess would be correct.

For the atom, the nucleus and the proton, then, the mass of the system is at least as large as the kinetic energy of the constituents and in some cases is much larger. If quarks and leptons are composite, however, the relation of energy to mass must be quite different. Since the prequarks have energies well above 100 GeV, one would guess that they would form composites with masses of hundreds of GeV or more. Actually the known quarks and leptons have masses that are much smaller; in the case of the electron and the neutrinos the mass is smaller by at least six orders of magnitude. The whole is much less than the sum of its parts.

The high energy of the prequarks is also what spoils the idea of viewing the higher generations of quarks and leptons as excited states of the same set of prequarks that form the first-generation particles. As in the other composite systems, the energy needed to change the orbits of the prequarks should be of the same order of magnitude as the kinetic energy of the constituents. One would therefore expect the successive generations to differ in mass by hundreds of GeV, whereas the actual mass differences are as small as .1 GeV.

At this point one might well adopt the view that the energy mismatch cannot be accepted, indeed that it simply demonstrates the elementary and structureless nature of the quarks and leptons. Many physicists hold this view. The energy mismatch, however, contradicts no basic law of physics, and I would argue that the circumstantial evidence for quark and lepton compositeness is sufficiently persuasive to warrant further investigation.

What is peculiar about the quark and lepton masses is not merely that they are

small but that they are virtually zero when measured on the energy scale defined by their constituents' energy. In other composite systems a small amount of mass is "lost" by being converted into the binding energy of the system. The total mass of a hydrogen atom, for example, is slightly less than that of an isolated proton and electron; the difference is equal to the binding energy. In a nucleus this "mass defect" can reach a few percent of the total mass. In a quark or a lepton, it seems, the entire mass of the system is canceled almost exactly. Such a "miraculous" cancellation is certainly not impossible, but it seems most unlikely to happen by accident. Similar large cancellations are known elsewhere in physics, and they have always been found to result from some symmetry principle or conservation law. If there is to be any hope of constructing a theory of prequark dynamics, it is essential to find such a symmetry in this case.

There is a likely candidate: chiral symmetry, or chirality. The name is derived from the Greek word for hand, and the symmetry has to do with handedness, the property defined by a particle's spin and direction of motion. Like other symmetries of nature, chiral symmetry has a conservation law associated with it, which gives the clearest account of what the symmetry means. The law states that the total number of right-handed particles and the total number of left-handed ones can never change.

In the ordinary world of protons, electrons and similar particles handedness or chirality clearly is not conserved. A violation of the conservation law can be demonstrated by a simple thought experiment. Imagine that an observer is moving in a straight line when he is overtaken by an electron. As the electron recedes from him he notes that its spin and direction of motion are related by a right-hand rule. Now suppose the observer speeds up, so that he is overtaking the electron. In the observer's frame of reference the electron seems to be approaching; in other words, it has reversed direction. Because its spin has not changed, however, it has become a left-handed particle.

There is one kind of particle to which this thought experiment cannot be applied: a massless particle. Because a massless particle must always move with the speed of light, no observer can ever go faster. As a result the handedness of a massless particle is an invariant property, independent of the observer's frame of reference. Furthermore, it can be shown that none of the known forces of nature (those mediated by the photon, the gluons and the weak bosons) can alter the handedness of a particle. Thus if the world were made up exclusively of massless particles, the world could be said to have chiral symmetry.

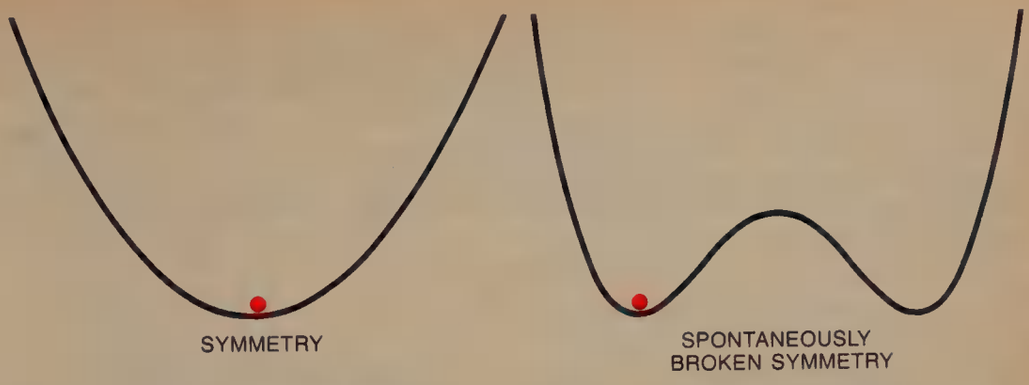
Chiral symmetry is the root of an idea that might conceivably account for the small mass of the quarks and leptons. The argument runs as follows. If the prequarks are massless particles, if they have a spin of 1/2 and if they interact with one another only through the exchange of gauge bosons, any theory describing their motion is guaranteed to have a chiral symmetry. If the massless prequarks then bind together to form composite spin-1/2 objects (namely the quarks and leptons), the chiral symmetry might ensure that the composite particles also remain massless compared with the huge energy of the prequarks inside them. Hence the small mass of the quarks and leptons is not an accident. They must be essentially massless with respect to the energy of their constituents if the chiral symmetry of the theory is to be maintained.

The crucial step in this argument is the one extending the chiral symmetry from a world of massless prequarks to one made up of composite quarks and leptons. It is essential that the symmetry of the original physical system survive in and be respected by the composite states formed out of the massless constituents. It may seem self-evident that if a theory is symmetrical in some sense, the physical systems described by the theory must exhibit that symmetry; actually, however, the spontaneous breaking of symmetries is commonplace. A familiar example is the roulette wheel. A physical theory of the roulette wheel would show it is completely symmetrical in the sense that each slot is equivalent to any other slot. The physical system formed by putting a ball in the roulette wheel, however, is decidedly asymmetrical: the ball invariably comes to rest in just one slot.

In the standard model it is the spontaneous breaking of a symmetry that makes the three weak bosons massive and leaves the photon massless. The theory that describes these gauge bosons is symmetrical, and in it the four bosons are essentially indistinguishable, but because of the symmetry breaking the physical states actually observed are quite different. Chiral symmetries are notoriously susceptible to symmetry breaking. Whether the chiral symmetry of prequarks breaks or not when the prequarks form composite objects can be determined only with a detailed understanding of the forces acting on the prequarks. For now that understanding does not exist. In certain models it can be shown that a chiral symmetry does exist but is definitely broken. No one has yet succeeded in constructing a composite model of quarks and leptons in which a chiral symmetry is known to remain unbroken. Neither the preon model nor the rishon model succeeds in solving the problem. The task is probably the most difficult one facing those attempting to demonstrate that quarks and leptons are composite.

If a consistent prequark theory can be worked out, it will still have to pass the test of experiment. First, it is important to establish in the laboratory whether or not quarks and leptons have any internal structure at all. If they do, experiments might then begin to discriminate among the various models. The experiments will have to penetrate the unknown realm of distances smaller than 10^{-16} centimeter and energies higher than 100 GeV. There are two basic ways to explore this region: by doing experiments with particles accelerated to very high energy and by making precise measurements of low-energy quantities that depend on the physics of events at very small distances.

Experiments of the first kind include the investigation of the weak bosons and



SPONTANEOUS SYMMETRY BREAKING is a mechanism that could spoil a prequark theory even if it has a chiral symmetry. Both of the physical systems shown here—a simple trough and a trough with a bump in the bottom—can be described as symmetrical in the sense that exchanging left and right leaves the system unaltered. For the simple trough the system remains symmetrical when a ball is put in the trough; the ball comes to rest in the center, so that exchanging left and right still has no effect. In the trough with a bump, however, the ball takes up a position on one side or the other, and the symmetry is inevitably broken. Similarly, a prequark theory that has a chiral symmetry might nonetheless give rise to composite systems that do not observe the symmetry. Showing that a chiral symmetry can definitely remain unbroken is currently the principal challenge in formulating a theory of how prequarks move.

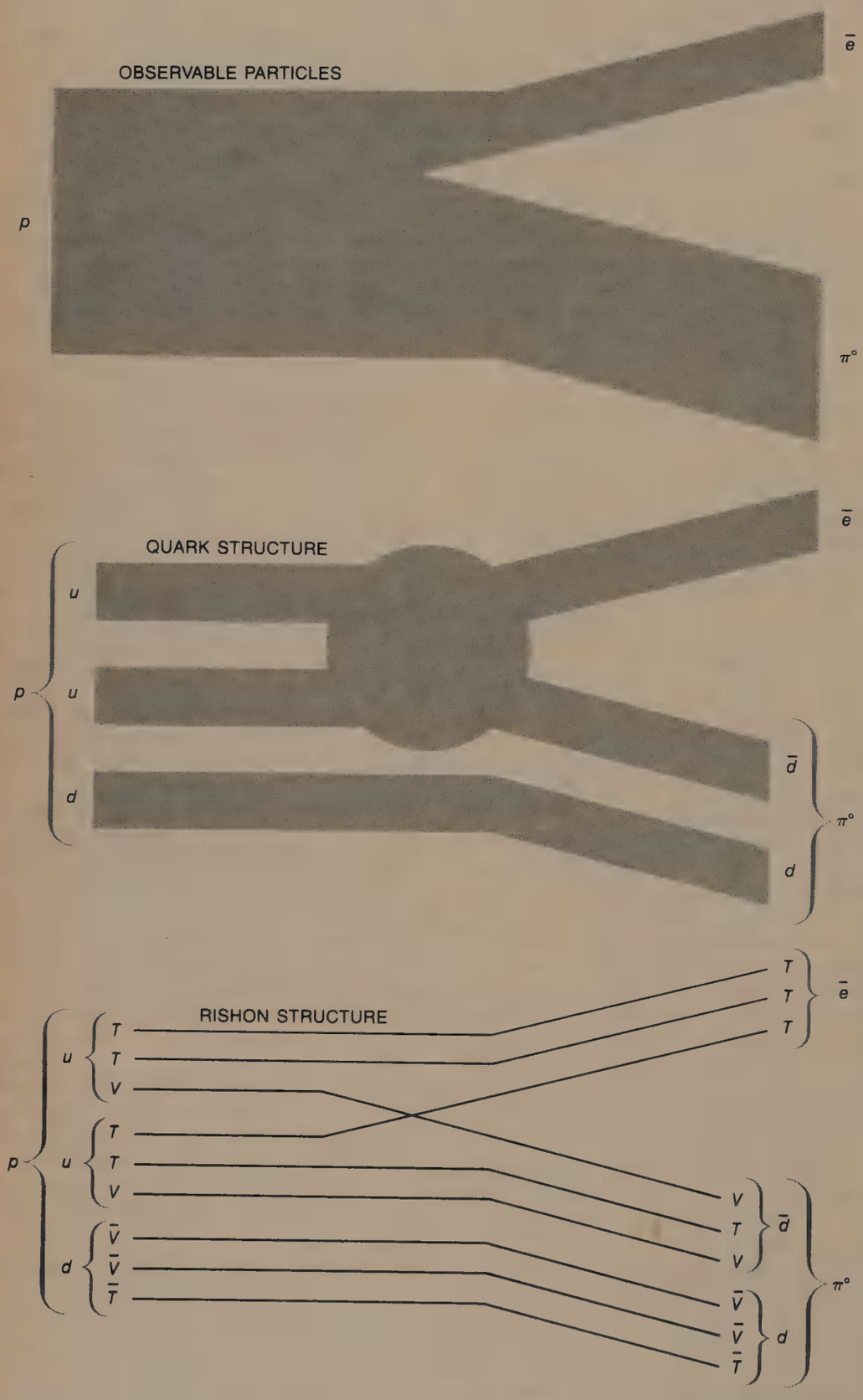
the search for the Higgs particles of the standard model. When such particles can be made in sufficient numbers, a careful look at their properties should reveal much about the physics of very small distances. New accelerators now being planned or built in the U.S., Europe and Japan are expected to yield detailed information about the weak bosons and will also continue the ongoing investigation of the quarks and leptons themselves.

Equally interesting are the high-precision, low-energy experiments. One of these is the search for the decay of the proton, a particle that is known to have an average lifetime of at least 10^{30} years. Several experiments are now monitoring large quantities of matter, incorporating substantially more than 10^{30} protons, in an attempt to detect the signals emitted when a proton disintegrates. None of the forces of the standard model can induce such an event, but none of the rules of the standard model absolutely forbids it. Both the grand unified theories and the prequark models, on the other hand, include mechanisms that could convert a proton into other particles that would ultimately leave behind only leptons and photons. If the decay is detected, its rate and the pattern of decay products could offer an important glimpse beyond the standard model.

There is similar interest in the hypothetical process in which a muon emits a photon and is thereby converted into an electron. Again none of the forces of the standard model can bring about an event of this kind, but again too no fundamental law forbids it. Some of the composite models allow the transition and others do not, so that a search for the process might offer a means of choosing among the models. Experiments done up to now put a limit of less than one in 10 billion on the probability that any given muon will decay in this way. Detection of such events and a determination of their rate might illuminate the mysterious distinction between the generations.

A third class of precision experiments are those that continue to refine the measurement of the magnetic moment of the electron and of the muon. Further improvements can be expected both in experimental accuracy and in the associated calculations of quantum electrodynamics. If the results continue to agree with the predictions of the standard model, the limit on the possible size of any quark and lepton substructure will become remoter. If a discrepancy between theory and experiment is detected, it will represent a strong hint that quarks and leptons are not elementary.

It may well be a decade or two before the next level in the structure of matter comes clearly into view (if, again, there is another level). What is needed is a sound theoretical picture, one that is self-consistent, that agrees with all experiments and that is simple enough to explain all the features of the standard model in terms of a few principles and a few fundamental particles and forces. The correct picture, whether it is a grand unified theory or a composite model of the quarks and leptons, may already exist in some embryonic form. On the other hand, it is also possible the correct theory will emerge only from some totally new idea. In the words of Niels Bohr, it may be that our present ideas "are not sufficiently crazy to be correct."



DECAY OF THE PROTON is a conjectured event that might be interpreted as experimental evidence for a grand unified theory or for a model of quark substructure. In one form of decay the proton would be observed to disintegrate into a positron ($\bar{e}$) and a neutral pion (π^0). The event can be understood in terms of the proton's quark constituents: an interaction of the two *u* quarks converts one of them into a positron and the other into a $\bar{d}$ antiquark; the latter combines with the remaining *d* quark of the proton to form the neutral pion. Grand unified theories suggest that the interaction of the *u* quarks is mediated by a new force of nature. The rishon model provides an alternative explanation: the two *u* quarks merely exchange a *T* and a *V* rishon.

Reshaping the future of film.

From fat to flat. The goal many of us strive for with laborious push-ups and strenuous diets, scientists have achieved in Kodak laboratories. The "bodies" they have streamlined, however, are those of silver halide grains. More than a beauty treatment, this process of reshaping enabled us to position grains for maximum light-catching ability, and expand their sensitizing characteristics. It's this advance in silver halide emulsion technology which made it possible for us to produce a true 1000-speed film with a balance of speed, grain, and sharpness previously unattainable.

The structure of new Kodacolor VR 1000 film is a modification of the inverted-layer format first used in Kodacolor 400 film. This concept places the fastest light-sensitive layers closer to the camera lens. The magenta and cyan layers are repositioned to form a fast-magenta, fast-cyan, slow-magenta, slow-cyan sequence. In the new emulsion, the yellow-filter layer is eliminated, and the protective overcoat and ultraviolet-filter layers are combined into one, without any loss in color quality.

The resultant faster film can be exposed by daylight, electronic flash, or tungsten, which provides photographers with greater picture-taking capability, particularly when recording fast-action scenes or shooting in low-light conditions.

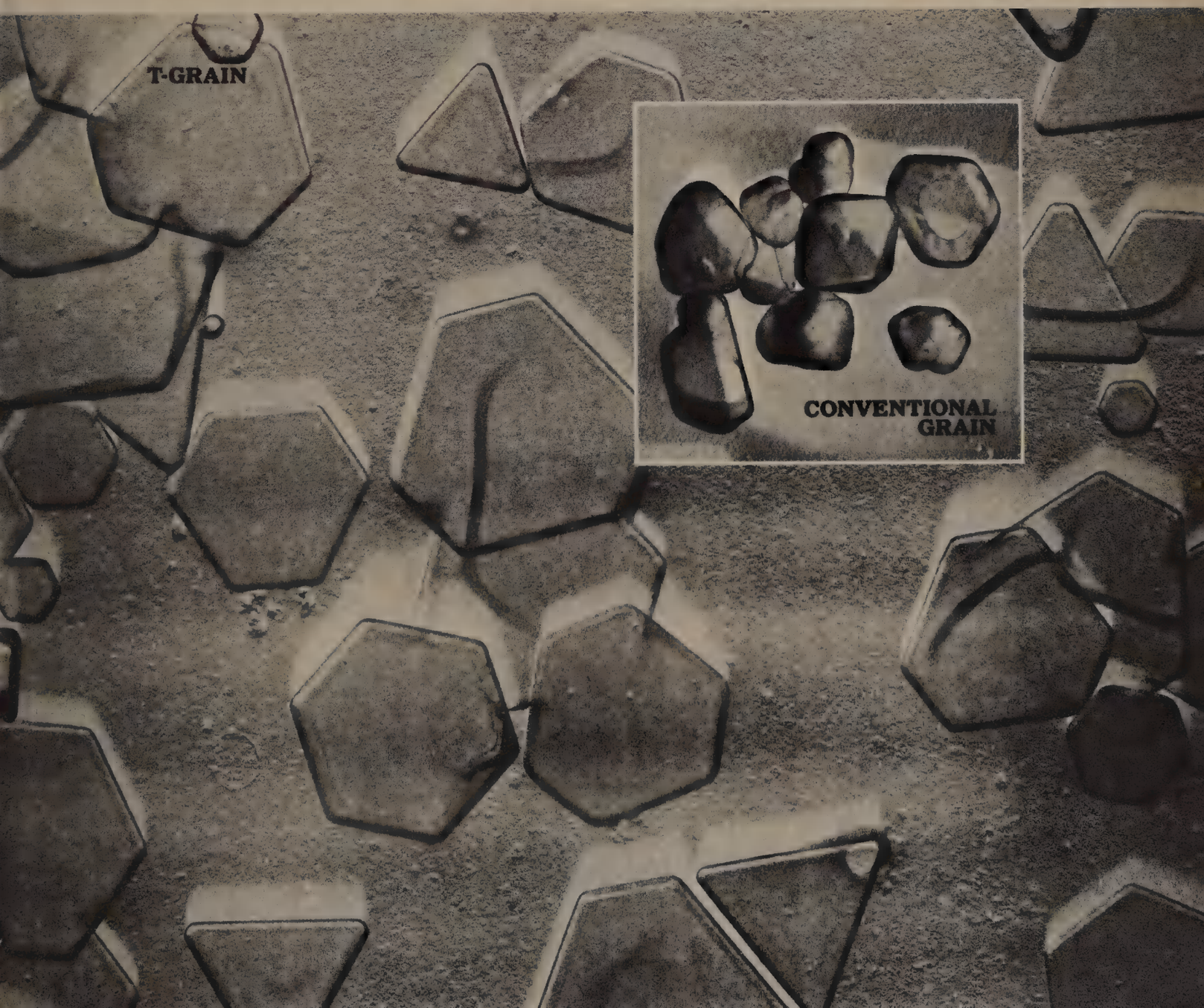
While reshaping the light-sensitive substance of film is a significant accomplishment in itself, we view this new technology as a stepping-stone. It may well reshape the future of film and enable photography to do hundreds of different jobs better than before.

For more information on Kodak's fastest color print film ever and the new emulsion technology behind it...send your request to: Eastman Kodak Company, Department GBSA-8, 343 State St., Rochester, NY 14650.

© Eastman Kodak Company, 1983

Kodak

Kodak. Where technology anticipates need.



SCIENCE AND THE CITIZEN

A Social Price

Amounts spent on arms and armies continued to increase last year, reaching a total of $600 billion. The total is given by Ruth Leger Sivard in "World Military and Social Expenditures 1982," the latest in a series of reports begun 15 years ago by the U.S. Arms Control and Disarmament Agency and now privately sponsored. The report aims "to provide an annual accounting of the use of world resources for social and for military purposes, and an objective basis for assessing relative priorities." The contention of the report is that "relative to the economy, the burden [of military expenditures] is growing larger." In other words, the world is paying a social price for the arms race.

From 1960 through 1980 world military expenditures increased by 441 percent. If the amounts are converted into 1979 dollars to remove the effects of inflation, the increase is 66 percent. Over the same period the sum of the gross national products of all countries (adjusted for inflation) increased by 146 percent. The average annual increase in military spending was 3.3 percent, and the average annual increase in G.N.P. was 7.3 percent. During most of the past 20 years, therefore, armaments claimed a decreasing share of the world's wealth. Between 1979 and 1980, however, military spending corrected for inflation went up by 3.6 percent and the average G.N.P. by only 2.6 percent.

The perspective is slightly different if increases in population, which offset part of the gain in G.N.P., are taken into account. The per capita G.N.P. calculated in 1979 dollars increased by 68 percent in the years from 1960 through 1980, or an average of 3.4 percent per year, much less than the average 7.3 percent increase in total G.N.P. Between 1979 and 1980 the per capita G.N.P., corrected for inflation, rose by less than 1 percent, compared with the 3.6 percent increase in military spending.

When the nations of the world are broken down into the two categories of developed and developing countries, it becomes apparent that the developing countries have paid a comparatively heavy social price for military spending. The military expenditures of the developed countries (in 1979 dollars) have increased by 44 percent in the past 20 years, whereas those of the developing countries have increased by 279 percent. The population of the developed countries increased by 20 percent and that of the developing countries by 58 percent. Because of the difference in population growth, the increases in per capita G.N.P. were 102 percent and 91 percent respectively, even though the G.N.P. of the developing countries increased by 219 percent and that of the developed countries by 130 percent.

The U.S. and the U.S.S.R. spend half of the world's military budget. "Both countries," the report comments, "the U.S.S.R. especially, have used a larger-than-average share of an expanding economic base to support the military effort." Whereas 4.5 percent of the world G.N.P. went for military expenditures in 1979, the U.S. spent 5.2 of its G.N.P. on the military and the U.S.S.R. spent 10.7 percent (although there are uncertainties in estimates of Russian military spending). The U.S. and the U.S.S.R. are also currently investing at least twice as much in military research and development as they are in nonmilitary research and development.

The report concludes with a breakdown of the military and social expenditures of 141 nations in 1979. Each nation is assigned a social-economic rank based on per capita G.N.P. and various indicators of the quality of education and health care, such as the population per physician. The U.S. is ranked seventh, after the Scandinavian countries, France and Australia. The U.S.S.R. is 23rd on the list.

W

With the discovery of the W particle a major goal of experimental elementary-particle physics has been achieved. The W belongs to the class of particles known as intermediate vector bosons, or weakons. According to the now-standard theory linking electromagnetism and the weak nuclear force, there are three such particles: two of them are electrically charged (the W^+ and W^-) and one is neutral (Z^0). In the unified electroweak theory the bosons act as intermediaries between weakly interacting particles, just as the photon serves as the carrier of the electromagnetic force between electrically charged particles. In contrast to the massless photon, however, the W and Z bosons have a prodigious mass: more than 80 times that of the proton or the neutron. Indeed, the large mass of the weakons long put them out of reach of particle accelerators [see "The Search for Intermediate Vector Bosons," by David B. Cline, Carlo Rubbia and Simon van der Meer; SCIENTIFIC AMERICAN, March, 1982].

The first instrument with enough energy to create such particles was built almost two years ago by modifying an accelerator at CERN, the European Laboratory for Particle Physics in Geneva. The CERN machine, which was designed primarily to look for intermediate vector bosons, brings counterrotating beams of protons and antiprotons into collision at a total energy of 540 GeV (billions of electron volts). After a series of preliminary runs in 1981, followed by an unexpected delay caused by an accident in one of the accelerator's two main particle detectors, the search resumed in earnest late last year. During a 30-day period in November and December approximately a billion high-energy proton-antiproton collisions were recorded for subsequent computer-aided analysis. By January physicists at CERN were able to announce that they had at last detected evidence of the fleeting existence of W particles in the aftermath of a small fraction of the collision events.

Two experimental groups were responsible for the investigation. One group, headed by Carlo Rubbia of CERN and Harvard University, included 126 physicists from 13 universities and research centers in Europe and the U.S. The other, headed by Pierre Darriulat of CERN, numbered 51 physicists from six European laboratories.

The first group, working with a large, multipurpose particle detector named UA-1 (for Underground Area One), has so far identified at least five events "revealing the expected signature of the W boson." According to an account of the group's findings submitted to *Physics Letters,* two independent analytic techniques were employed to distinguish the events of interest from the large number of other events. In one approach the group looked for the track of a high-energy electron or positron (the electron's antiparticle) emitted from the collision zone at a large angle with respect to the axis of the colliding beams. In the other approach the group looked for evidence of a significant amount of "missing energy" attributable to the emission of a neutrino from the collision zone. (Because neutrinos have no electric charge, they cannot be tracked by the electromagnetic detection system.)

In each case the experimenters were left with five unambiguous W candidates and one probable candidate. Comparison of the two sets of data showed that both methods of analysis had identified the same events. As the experimenters note in their report, "the simultaneous presence of an electron and (one) neutrino of approximately equal and opposite momenta in the transverse direction... suggests the presence of a two-body decay." The decay is interpreted as the spontaneous disintegration of a short-lived W^- particle into an electron and an antineutrino or of an equally short-lived W^+ particle into a pos-

DATSUN MAXIMA
OBVIOUS LUXURY...SURPRISING PERFORMANCE

Introducing the luxury sedan that was born to perform. The magnificent Datsun Maxima boasts a fuel-injected 2.4 liter six-cylinder engine that merges surprising power with unsurpassed elegance.

Inside, virtually every conceivable creature comfort. cruise control on both the 5-speed and automatic, eight speaker stereo including cassette deck with Dolby, air conditioning, power windows, mirrors and door locks. A voice will tell you when your "Right door is open"—one of six vocal reminders. All this is standard.

The optional sunroof is power driven. The optional leather package includes a new digital control panel that monitors the engine's functions electronically.

The Datsun Maxima is available in four-door sedan or wagon, gas or diesel. Compare it feature by feature with any other car in its luxury class. Compare the performance. Then compare the price. Most importantly of all, remember it's built by the Nissan Motor Company, Ltd.—whose name stands for quality world-wide.

DATSUN
WE ARE DRIVEN

THE NAME OF QUALITY
NISSAN

Product of NISSAN

1983 NISSAN MOTOR CORPORATION IN USA

The world's densest compute is now the heart of the world's mos 32-bit computer.

Our 32-bit CPU.

From time to time, miracles of technology come along to make previously impossible tasks not only possible, but easy. That tiny 450,000-transistor integrated circuit is one of those technological miracles.

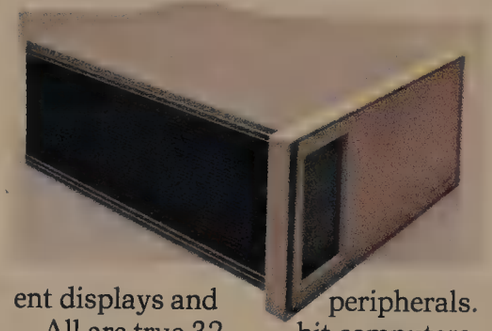
Hewlett-Packard didn't develop it just to break the record for most transistors on a chip, but to put on an engineer's or scientist's desk a computer so powerful that it can do the work of mainframes costing four times as much.

32-bit computers for 32-bit applications.

The new HP 9000 computer based on this and four other 'superchips' can handle formidable engineering and scientific problems. The scientist solving complex systems of equations, the mechanical engineer doing finite element analysis or three-dimensional modeling, the electrical engineer analyzing complex circuits or designing very large-scale integrated circuits — these are the kinds of technical people and problems the HP 9000 family is designed for.

It comes in three versions. The integrated workstation is complete with keyboard, color or monochromatic graphics display, fixed and flexible disc drives, and printer. For systems manufacturers, there's a rack-mountable box. And for a variety of single-user and multi-user applications, the minicabinet version works with many different displays and peripherals.

A rack-mountable version is available, too.



All are true 32-bit computers, with 32-bit CPUs, memories, and data paths. And the multi-CPU architecture lets you nearly double or triple your processing power at any time by adding one or two CPU boards. Without increasing the computer's size.

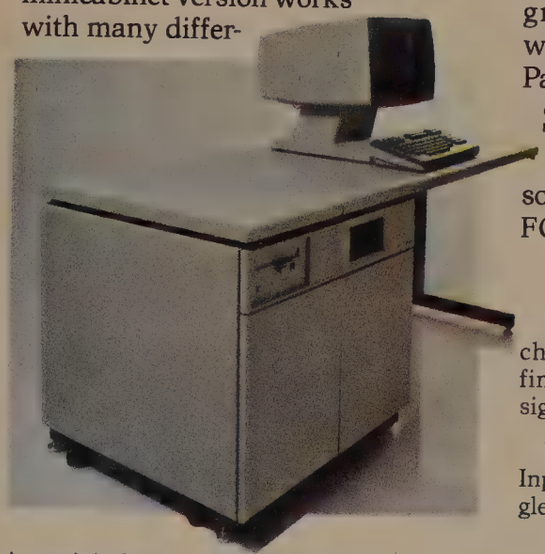
Two operating systems are better than one.

The integrated workstation is available with a choice of operating systems. One is HP's highly evolved, high-performance Enhanced BASIC, augmented with 3-D graphics and a software innovation called a run-time compiler. This substantially increases program execution speed, while retaining an interactive development environment.

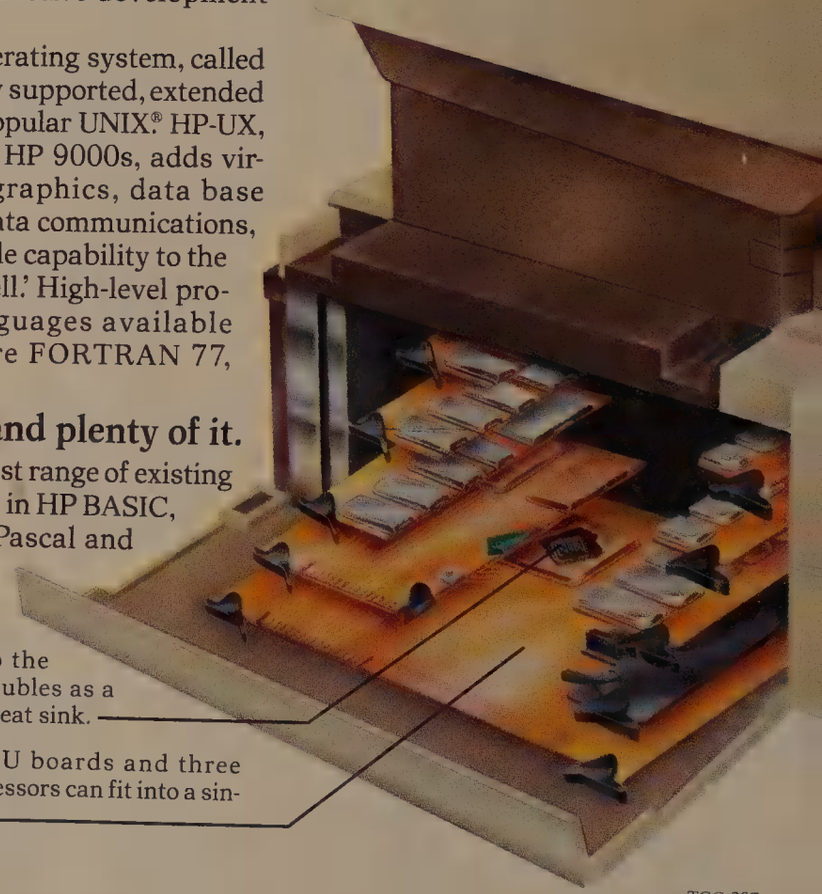
The other operating system, called HP-UX, is a fully supported, extended version of the popular UNIX® HP-UX, available on all HP 9000s, adds virtual memory, graphics, data base management, data communications, and enhanced file capability to the basic UNIX 'shell.' High-level programming languages available with HP-UX are FORTRAN 77, Pascal and C.

Software, and plenty of it.

Much of the vast range of existing software written in HP BASIC, FORTRAN 77, Pascal and C is transportable to the HP 9000. HP will also be offering proprietary software packages emphasizing computer-aided design and engineering. These will tie the HP 9000 into HP's Manufacturer's Productivity Network (MPN). Third-party software suppliers will be providing many of the most widely used CAE packages for 32-bit computer systems. And both HP 9000 operating



As a minicabinet, it can handle multiple users.



The 32-bit CPU chip is bonded to the finstrate which doubles as a signal carrier and heat sink.

Up to three CPU boards and three Input/Output Processors can fit into a single HP 9000.

TCG-207

hip
ffordable

systems offer extensive program development tools.

You also get a choice of communication tools. The HP 9000 is currently compatible with Ethernet,™ and with HP's Shared Resource Manager (SRM) which lets clusters of HP 9000 and 16-bit desktop computers share data and use common peripherals.

Links to central computers are also available. And in late 1983, HP will offer local area networks based on the IEEE-802 standard.

New technology from the silicon up.

The five superchips that make the HP 9000 possible are the 32-bit CPU, which can execute a million instructions per second; an eight-channel Input/Output processor (IOP); a random-access memory chip capable of storing 128K bits of data; a memory controller that 'heals' up to 32 bad memory locations; and an 18-megahertz clock.

Hewlett-Packard's advanced NMOS-III process makes it possible to put 450,000 transistors on a chip only 0.4 centimeters square. This tremendous density of electronic components could have required an expensive and elaborate cooling system.

Instead, HP engineers developed a new mounting structure called a finstrate, a copper-cored circuit board, which acts as both cooling fin and substrate. The finstrates containing the CPU, IOP, memory, and clock chips are housed in a lunchpail-sized module.

One user, one mainframe.

Clearly the trend in engineering and scientific computation is away from large machines shared by multiple users and towards networks of powerful personal workstations, sharing peripherals and data bases. The reason is compelling. An engineer or scientist in personal control of an HP 9000 can solve so many more problems more easily that the increased productivity alone makes the cost of individual computers easy to justify.

For complete information about this powerful breakthrough in 32-bit computing, contact the local HP sales office listed in your telephone directory. Ask a Technical Computer Specialist for a demonstration. Or write to Pete Hamilton, Dept. 41151, Hewlett-Packard, 3404 East Harmony Road, Fort Collins, CO 80525.

HEWLETT PACKARD

UNIX is a trademark of Bell Laboratories. Ethernet is a trademark of Xerox Corporation.

The Crown Jewel of England.

itron and a neutrino. The mass of the W particles implied by measurements of the decay products is 81 ± 5 GeV, which the group points out is "in excellent agreement" with the prediction of the electroweak theory.

The second group, working with a more specialized and somewhat simpler detector named UA-2, announced in January that it had found four events "that are consistent with a W signature." The group added, however, that "more work is needed to confirm this preliminary UA-2 result." Both groups will get a chance to substantiate their early findings when the CERN proton-antiproton collider starts a new series of observations this month. High on the agenda will be the search for the Z^0, which ought to materialize about 10 times less frequently than the W particles.

Unpolishing the Apple

In the academic year 1980–81 there were about 140,000 students at the nine campuses of the University of California. Of this number 22 students were enrolled in programs leading to certification as a teacher of mathematics in elementary or secondary school. There were 47 students in courses leading to certification as a science teacher. In the same academic year there were about 300,000 students attending the 19 campuses of the California State University system; 75 were in training to be mathematics teachers and 127 to be science teachers. The numbers are cited by James Guthrie and Ami Zusman of the University of California School of Education in a report on the training of mathematics and science teachers in California; the report was published recently by the university's Institute of Governmental Studies.

One reason few students seem to be interested in becoming mathematics or science teachers is the inadequacy of teachers' salaries. The average salary for a beginning teacher in the San Francisco Bay area is $12,680. University graduates with the qualifications needed to become a certified teacher can get jobs in industry at a starting salary of about $20,000. In addition to reducing the number of future teachers, the higher pay offered by industry is depleting the current ranks of elementary and high school teachers. Guthrie and Zusman cite a survey done in the high school districts around Palo Alto. The area is part of the region known as Silicon Valley and has a high concentration of businesses involved with advanced technology. The survey found that six of the eight school districts in the area were losing mathematics and science teachers to better-paid jobs in industry.

Ironically, it is in areas where industry is based largely on new technology that the need for excellence in science educa-

tion is most acute. Jobs in such industries require a comparatively high level of scientific and technical knowledge. Guthrie and Zusman estimate that between 1980 and 1990 about 40 percent of the new jobs in California will be in businesses that employ new technology. Thus California is faced with an increasing demand for technically trained workers and a decreasing supply of teachers to educate them.

The disjunction between supply and demand in science education may be greater in California than it is elsewhere in the U.S., but the problem is of national scope. In 1982 the National Science Board of the National Science Foundation established a Commission on Precollege Education in Mathematics, Science and Technology. In its first formal report the commission concludes that the national system of science education is failing to give workers adequate technical skills. Indeed, the commission concludes that this is one of two major functions of science education that are not being fulfilled. The other is providing the general public with sufficient knowledge to enable citizens to make informed decisions on matters related to science and technology. A third function, that of training professional scientists, mathematicians and engineers, is being carried out better than the other two, the commission states.

The shortage of qualified teachers was found to be widespread. In 1981, 43 states out of the 45 that responded to a survey said they had a shortage of mathematics teachers; 42 of the 45 had a shortage of physics teachers. From 1971 through 1980 the number of students preparing to become teachers decreased by a factor of three in science and by a factor of four in mathematics. Moreover, of those who had trained to teach between 1971 and 1980 only half actually entered the teaching profession. A fourth of those who were teaching said they planned to leave for other jobs in the near future.

In an attempt to reverse the trend the Reagan Administration has proposed appropriating $200 million for the training of mathematics and science teachers. The funds would enable states to make grants of up to $5,000 to teachers who want to take up mathematics or science teaching. Those eligible for the grants might include unemployed teachers, teachers in other fields of study and teachers of mathematics or science who have no certification in those subjects.

According to the Commission on Precollege Education, only a minority of high school students in the U.S. now receive an adequate scientific education. Only a third of the 21,000 American high schools offer a course in calculus. Fewer than a third offer a course in physics taught by a qualified teacher. The report concludes: "The educational system may actually have carried out [the education of skilled workers and the general public] better 20 years ago: the proportion of public high school students enrolled in science courses has declined since that time."

Oncogenetic Switch

Human cells harbor genes that have a function in normal metabolism but that can be triggered to promote cancerous growth. The discovery of such oncogenes has provided both a conceptual framework and a set of powerful tools for investigating the molecular genetics of cancer. A number of lines of research have now begun to make sense of what had seemed to be disparate observations concerning the causes and manifestations of malignancy.

The essential characteristic of a cancer is the uncontrolled proliferation of cells. The product of one family of oncogenes is a protein kinase, an enzyme that adds phosphate groups to proteins; the oncogene kinases are unusual in that they phosphorylate the amino acid tyrosine, whereas most kinases phosphorylate serine or threonine. The receptors for insulin and some other cellular growth factors can also phosphorylate tyrosine. The receptors do so, however, only in the presence of the appropriate growth factor, whereas the oncogene enzymes throw a switch that commits cells to uncontrolled proliferation. The phosphorylation of tyrosine may therefore be central to growth control, and aberrant phosphorylation may be one step in carcinogenesis.

Identifying the proteins encoded by oncogenes and determining what the proteins do in the cell is one objective of current research. Another compelling objective is to learn what activates an oncogene, changing a normal gene into a cancer gene. Perhaps the carcinogenic version of the gene has undergone mutation or reshuffling, so that it specifies a different protein: an enzyme or regulatory molecule that is either more or less efficient than the normal one, say, or an enzyme with an altered substrate specificity. Perhaps, on the other hand, the protein-coding region of the gene is not changed but the level of its expression is enhanced, so that too much of a normally benign protein is synthesized. In the past few months some striking results from two laboratories have pinpointed a possible mechanism of oncogene activation that would be compatible with either the mutational or the dosage hypothesis or with both.

A number of cancers, particularly those affecting various blood and lymphatic cells, are associated with chromosomal abnormalities. One kind of abnormality is a reciprocal translocation, in which segments of two chromosomes are interchanged. Several investigators, including John Cairns of the Harvard School of Public Health, George Klein of the Karolinska Institute and Janet D. Rowley of the University of Chicago School of Medicine, have pointed out that a translocation might move a potential oncogene into a region where DNA is being actively transcribed and thereby enhance its expression. Last year several oncogenes were mapped to specific chromosomal segments that are indeed subject to translocation.

The gene designated c-*myc* is the cellular analogue of a viral oncogene originally recognized for its ability to induce tumors in chickens and other birds. The gene is near one end of human chromosome 8. In a chromosomal abnormality associated with Burkitt's lymphoma, a cancer of antibody-producing *B* lymphocytes, the end of chromosome 8 is translocated. Most frequently it moves to chromosome 14, to a region carrying genes for the heavy chain of antibody molecules.

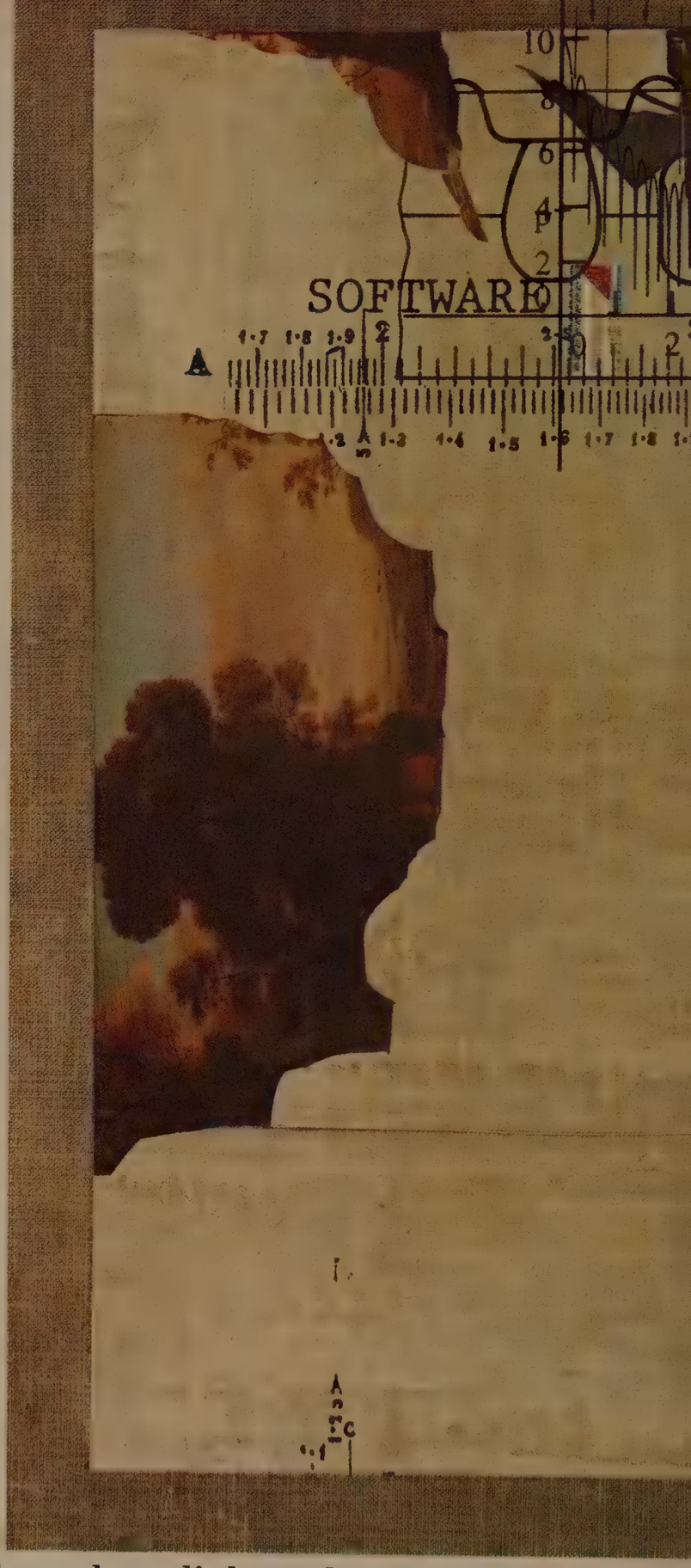
In lymphocytes the antibody-gene loci are particularly active ones; in the course of the immune response they are shuffled in various ways and are transcribed into RNA that is translated to make specific antibodies of various classes. Now two groups, headed by Philip Leder of the Harvard Medical School and by Carlo M. Croce of the Wistar Institute of Anatomy and Biology, have reported in *Proceedings of the National Academy of Sciences* that the locus of c-*myc* on chromosome 8 is precisely at the "breakpoint" of the 8-to-14 translocation. In several Burkitt's-lymphoma cell lines c-*myc* apparently ends up associated with a particular heavy-chain antibody gene on chromosome 14.

The implication is that translocation to the heavy-chain locus brings c-*myc* under the influence of a "promoter" region that ordinarily mediates the transcription of a particular heavy-chain gene; in Burkitt's lymphoma the region may instead promote the active transcription of c-*myc* and thereby enhance its expression. (So far Leder's group has not observed a dramatic increase in c-*myc* RNA in lymphoma cell lines; Croce's group reports a significant increase.) An alternative possibility is that in the course of the translocation c-*myc* is somehow rearranged so that it encodes a different product. Further analysis of the translocated gene, together with more study of c-*myc*'s protein product (of which little is yet known), should tell whether mutation or dosage is at work in this case.

Spectral Memory

The presence or absence of a mark at a predetermined position on a sheet of paper encodes one bit of information. If the mark is allowed to vary in color as well as position, the density of the infor-

What do IBM people think about?



Creating better computers takes a lot of different thinking from a lot of different people.

At IBM, our scientists and engineers represent a wide range of disciplines.

For example, IBM chemists are exploring the structure of ceramics and developing special polymers to make computers more compact.

IBM physicists are studying how short light pulses travel through optical fibers to build faster computer systems.

There are IBM psychologists and human factors specialists who study how people interact with machines, so we can make them friendlier and easier to use.

In laboratories and plants all over the world, thousands of

IBM people are putting their minds to work on questions of science and technology. They're all part of a 50-year commitment to research and development—a continuing commitment that has totalled more than $8 billion in the last six years alone.

Because finding new ways to improve our computers is one thing that's always on our minds.

IBM®

Some of the most powerful forces in communications are at work at CSD

Systems Engineers
Hardware Engineers
Software Engineers
ATE Engineers
R/M Engineers

We want engineers who can see the total picture, whether they're working on independent R&D or leading an entire department. People who can translate customer needs into creative solutions, identify technologies, make the right design decisions and lead in high visibility positions are going to succeed here.

Engineers like you have found unlimited growth at CSD. Some of the leading minds in communications engineering are working here today on projects like those described in this ad. Our involvement goes across the entire communications spectrum from DC to daylight.

We are now growing so rapidly that **if you meet the criteria described in this ad, we can create the right opportunity for you.**

Among our broad range of projects:

TRANSMISSION SYSTEMS

We're engineering integrated radio systems using digital wide band processing. New technologies include unique fiber optic applications and systems operating in the VHF to EHF range.

MILSATCOM

Our emergent-technology EHF band systems involve AJ protection, rapid configuration and interoperability with satellite networks. Key systems elements include EHF antennas, feeds and high performance synthesizers.

DIGITAL PROCESSING SYSTEMS

These high performance systems cover the range of audio, ELF, VLF, LF and HF. We're developing algorithms with critical timing and throughput requirements, as well as applying high speed data processing in advanced digital communications products that include digital modems, voice and receivers, and systems controllers.

NETWORKING

We're pioneering totally integrated multinetwork interface systems pushing the state-of-the-art. Our extensive involvement in networking includes new concepts in such areas as message, packet and digital switching.

SECURE SYSTEMS

The security/intelligence community looks to us for leadership in systems knowledge. We're providing custom LSI/VLSI, dedicated secure communications networks, fault tolerant computer architectures and computer network security.

SURVIVABLE and ENDURING SYSTEMS

With our primary focus on C^3I propagation and network analysis under benign and stressed environments, we're expanding modeling and assessment software tools to determine vulnerability, survivability, endurability and/or exploitation of C^3I systems worldwide.

Along with a network of project opportunities offering new growth comes a strong system of support and rewards. Success here is recognized within GTE and across the communications industry. Internal advancement can take the direction of management or development of technical knowledge. Our educational opportunities are unusually sophisticated and include an in-house MASTERS DEGREE program in Systems Engineering and Computer Science, and complete tuition reimbursement. Our benefits include life, health and dental insurance coverage for dependents. And because we provide year after year of challenge, CSD has kept top quality professionals and maintained low turnover in a fiercely competitive market for engineers. For access to the important opportunities in communications, we can be the last place you'll ever have to look.

We will also consider candidates who wish to enter Systems Engineering and have strong experience in our industry.

Send your resume, indicating areas of interest, to Greg M. Gostanian, GTE's Sylvania Systems Group, Communication Systems Division, 77 "A" Street, Needham, MA 02194. Or call us in Needham, MA at 1-800-225-3956.

An Equal Opportunity Employer M/F

If it's happening in communications, it's happening at CSD.

GTE Systems

mation that can be stored in a given area is multiplied by the number of colors that can be distinguished from one another at each site. An application of this principle is being studied by investigators at Cornell University, the International Business Machines Corporation, the General Motors Research Laboratories and Bell Laboratories. Instead of colored marks on paper the method exploits molecules that can effectively be given a color by laser irradiation. If the technique can be perfected, at least 1,000 colors might be distinguished at sites so small that 100 million of them could be packed into a square centimeter of surface area. Hence a storage density as high as 100 billion bits per square centimeter could be achieved.

A simpler method of optical recording for the bulk storage of information has already been demonstrated. A fine laser beam is directed onto a substrate and burns a small hole in the material. The presence or absence of a hole at a certain position encodes the information, and the size of the hole determines the storage density. The size is limited roughly by the wavelength of the laser radiation: the hole cannot be smaller than a square region one wavelength on a side. The technique now under study would combine the laser technology and a property of certain solids that was first described by Russian physicists in 1974 and has been investigated by workers at IBM since the late 1970's.

A solid can absorb radiation of a given frequency only if the molecules of the solid can resonate in response to the radiation and assume a higher energy state. In most circumstances the molecules return to the lower energy state as soon as the radiation is turned off; under some conditions, however, the molecules can undergo transformations that prevent them from falling back to their original state. Until recently it was thought that the molecular transformation must be a chemical one, in which the identity of the molecules is changed, but it has now been shown that the transformation can be a simple reorientation of the molecules. In both cases the effect is the same: the original, lower energy state is depopulated by the radiation, and the material ceases to resonate and absorb radiation at that frequency. (For visible radiation the frequency of the radiation corresponds to its color.) A persistent "hole" is created in the absorption spectrum of the material, and so the material becomes transparent across the narrow frequency band that corresponds to the resonant frequency of the original low-energy state.

Each hole in the absorption spectrum can correspond to one bit of information, and the spectrum available across each small region of a solid is wide enough to accommodate at least 1,000 spectral holes. All 1,000 bits can therefore be stored in an area one wavelength on a side. Investigation is now under way to enlarge the range of conditions and the family of materials that can be employed to make spectral holes.

Fire One

In man's reconstruction of his own prehistory an important development was the mastery of fire, which surely began with the capture and control of wildfire. In the 1930's evidence of hearths and charcoal dating to roughly 500,000 years ago was discovered at the *Homo erectus* site of Zhoukoudian, near Beijing. Hearths have since been found at Vértesszöllös in Hungary and at Tautavel in France, confirming that the knack of controlling fire was part of the behavioral repertory of *H. erectus.*

By the time of the European discoveries, however, *H. erectus* had been granted an antiquity substantially greater than 500,000 years. Remains of the species a full million years old had been found in the Far East and remains perhaps 1.5 million years old had been found in East Africa. If the species indeed had the cultural role of a Middle Paleolithic Prometheus (the Titan, it will be remembered, stole fire but did not make it), earlier evidence that *H. erectus* was able to control fire ought to exist.

The evidence may now have been found by a multidisciplinary team of investigators: J. W. K. Harris of the National Museums of Kenya, John Gowlett of the Research Laboratory for Archaeology and Art at the University of Oxford, Bernard Wood, an anatomist at the Middlesex Hospital Medical School in London, and Derek Walton of the department of physics at McMaster University. Working along the edge of the Rift Valley in the Laikipia Escarpment region of Kenya, they concentrated their efforts at a place called Chesowanja, on the east side of Lake Baringo, where artifacts and animal remains (including hominid remains) had already been found. The team soon discovered lumps of reddish mineralized matter. Walton, the geophysicist, determined that the lumps are composed of clay that has been baked to hardness at a temperature of 400 degrees Celsius. The distribution of the baked lumps indicated their direct association with stone artifacts, unmodified stone cobbles and fragmentary animal bones. At one site the baked material gave every appearance of being a manmade hearth.

Were the clay lumps formed by passing grass fires, a common event on the East African savanna? No; grass fires seldom raise the temperature of the ground above 65 degrees C. Might they be the result of exposure to more intense heat, such as smoldering tree stumps or strokes of lightning? Possibly; such chance events cannot be ruled out. Nevertheless, the association of the lumps of fired clay with hearth building, artifact manufacture and food consumption suggests that they are the result of human control of fire. How long ago was this? A chronology based on the radioactive decay of potassium into argon gives a date of about 1.4 million years ago. Who was tending the fire? Probably some early African example of *H. erectus.*

Coping with the Death Cap

Walter Litten, SCIENTIFIC AMERICAN's occasional correspondent on the "death cap" mushroom *Amanita phalloides* and other matters mycological, sends the following report:

"Even though poisoning by the death cap is not a bacterial disease and therefore seems unlikely to respond to antibiotics, evidence is accumulating that penicillin is a component of the most successful therapy for dealing with it. Two other components are oxygen-enriched air and silibinin, a substance from the milk thistle, *Silybum marianum.*

"A statistical analysis of 205 European cases of death-cap poisoning from 1971 through 1980 is presented in *Swiss Medical Weekly* by George L. Floersheim, O. Weber, P. Tschumi and M. Ulbrich. The three treatments were found to save more patients than were saved by any other therapy out of the 30 evaluated. The average number of therapies tried was eight, and some patients were subjected to as many as 20 therapies by physicians desperately 'trying everything.' Of the 205 patients, 45 died. This 22 percent mortality is no great advance over the 30 percent mortality recorded before the 1970's.

"Floersheim attributes the slight improvement not to better medical management but to greater awareness of the unreliability of traditional criteria for recognizing deadly species. He speculates that many of those who were poisoned ate sparingly of the mushrooms. A person who has swallowed much less of the toxins than is contained in one average-size specimen of *A. phalloides* has a good chance of recovery from the adventure in dining, given good supportive medical care.

"Floersheim states that the single most important factor distinguishing the survivors from the fatalities among the 205 patients is the quantity of *A. phalloides* ingested per unit of body weight. Analysis is complicated, however, by the considerable variation among mushrooms of the one species in their content of amatoxins, rings of eight amino acids that are considered principally responsible for the loss of liver function caused by the deadly amanitas. Recognizing these uncertainties, Floersheim and his colleagues nonetheless looked for a correlation between survival and the treatment administered.

THE FISH WON'T BE BITING IN TWITCHELL CREEK TODAY,

OR TOMORROW, EITHER.

Twitchell Creek on Woods Lake in the Adirondacks. Once a great fishing spot. But no more. The fish are gone now . . . killed off by the rain. Acid rain which poisoned the water, the land, and many of the creatures who once lived here.

Acid rain is deadly because it contains nitric acid and sulfuric acid. These killers are born when rain water mixes with nitrogen oxide and sulfur dioxide . . . two chemicals which are being spewed by the ton-load into our air every day from coal-burning power plants and industrial boilers.

Some lakes contain neutralizing "buffers" which lessen acid damage. But who will stop the rain in the Adirondacks, in western Virginia, throughout New England? The Izaak Walton League, that's who. Formed in 1922 by a handful of conservation-minded fishermen, the League is now 50,000 strong. And ready to fight those who want to relax the 1970 Clean Air regulations to allow for greater sulfur dioxide emissions. We want fewer emissions. Stricter laws. And we'll get them. Because there's something at stake here far greater than fishing. And that's life itself.

For more information on our activities, write:

THE IZAAK WALTON LEAGUE OF AMERICA

Izaak Walton League
1800 North Kent Street
Arlington, Virginia 22209

"Thioctic acid, a relatively simple organic molecule, has been advanced as an antidote to amatoxin, and arrangements have been made for rushing it to physicians treating severe cases. In the 205 cases studied thioctic acid therapy turned out to be one of the measures taken more often with patients who did not survive. In the same category were treatment with the drug sulfamethoxazole, treatment of hemorrhaging induced by the toxin, plasma expanders to compensate for a reduced volume of circulating blood, hemodialysis to try to remove the toxins from the circulation and buffering to counter metabolic acidosis. For 20 other therapeutic measures the analysis showed neither positive nor negative correlation with outcome. On the positive side, penicillin and hyperbaric oxygen were found to be independently associated with a higher survival rate. The rate was enhanced when treatment with silibinin was combined with the injection of penicillin.

"One of several hypotheses to explain the apparent effect of penicillin in death-cap poisoning has been suggested by D. F. Schaefer and E. A. Jones in *The Lancet*. The proposed mechanism centers on gamma-aminobutyric acid (GABA), the chief inhibitor of nerve transmission. Bacteria in the gut synthesize GABA, and in the liver it is metabolized. A liver crippled by the attack of amatoxin does an inadequate job of keeping the GABA level in check. The result may be neural inhibition and the severe encephalopathy that is likely to be the final cause of death from *A. phalloides* and other amatoxin-laden toadstools. The role of penicillin may be to suppress the intestinal flora and so prevent the buildup of GABA until the liver mends.

"One of the later developments in severe *A. phalloides* intoxication is slowed blood clotting as the liver fails to manufacture the clotting factor thromboplastin. Clotting time is measured in terms of a percentage known as the Quick value. In Floersheim's survey 84 percent of the patients who reached a Quick value below 10 percent died; of the patients whose Quick value dropped no lower than 40 percent, none died."

Playing It by Ear

For a baseball outfielder to catch a fly ball is one thing; to explain how he does it is quite another. In 1968 Seville Chapman of the Cornell Aeronautical Laboratory considered the fly-ball question and advanced the hypothesis that the outfielder subconsciously solves a problem in trigonometry: he runs to a position on the field where the tangent of the ball's angle of elevation increases at a constant rate. Peter J. Brancazio of Brooklyn College of the City University of New York has reexamined the matter and proposes that the key is not trigonometry but the fielder's inner-ear sensors of motion. "We may actually be judging fly balls by ear," Brancazio says in summarizing a paper he presented at the joint annual meeting of the American Physical Society and the American Association of Physics Teachers.

"One of the unique aspects of this skill," Brancazio writes, "is that there is no way to verbalize it or to teach someone how to do it. There do not seem to be any coaching techniques or helpful hints that could [improve] one's ability to judge fly balls accurately. The skill, it seems, must be completely self-taught on a nonverbal level."

In Chapman's view the outfielder teaches himself that if a ball is going over his head, he will see the tangent increasing at an increasing rate. (Brancazio calls this ball "Chapman's homer.") If the ball is going to fall in front of the fielder, he sees the tangent increasing at a decreasing rate. The fielder's art is one of learning where and how fast to run in order to keep the tangent increasing at a constant rate.

The difficulty with this hypothesis, Brancazio writes, is that it is based on the assumption that air resistance does not significantly affect the flight of the ball. "I have computed the effects of aerodynamic drag on the trajectory of a baseball and discovered that for the typical speeds and times of flight that occur under game conditions a batted baseball travels about 60 percent as far as it would in a vacuum. Moreover, air resistance distorts the shape of the trajectory of the ball so that it is noticeably different from a parabola. When aerodynamic forces are accounted for, the tangent of the angle of elevation does not increase at a constant rate as seen by a fielder who has judged the flight of the ball correctly."

Brancazio examined several other geometric and trigonometric features of the trajectory and found that none of them seems to show any characteristic variation that would tell the outfielder how to move to catch the ball. Considering the fly-ball problem as part of the larger question of how a person determines the position in space of any moving object and how he coordinates this information with the movement of his body, Brancazio notes that the person usually moves his head as well as his eyes as he follows the object. The motion of the head and that of the eyes must be closely coordinated. It turns out the coordination is guided primarily not by visual feedback but by signals triggered in the inner-ear sensors by the motions of the head. "This coordination of sensory input with body motion evidently follows a neural pathway that has been established through the familiar behavioral process of learning by trial and error."

WORKING FACTS:

- 3 percent of the U.S. labor force feeds the country and produces a surplus for export.
- After two centuries women's work still tends to be characterized by low pay and occupational segregation.
- Finance, distribution, and transport are mechanized even more than the production of goods.

THE MECHANIZATION OF WORK

A SCIENTIFIC AMERICAN BOOK

The easing of human labor by technology has spanned the development of the most elementary force-transmitting machines to those capable of the most advanced information handling. THE MECHANIZATION OF WORK is a collection of articles from the September 1982 issue of SCIENTIFIC AMERICAN that analyze the profound effects this phenomenon has had upon modern society, transforming the nature of work in ways never previously thought possible.

Focusing on changes in a wide variety of vocational spheres–from agriculture, mining, and design and manufacturing to commerce, office work, and "women's" occupations–this book explores the diversity of technologies and conditions contributing to development in these areas. Shifts in patterns of vocation selection are examined in detail, while concomitant issues of public policy and values are also considered.

Including a provocative discussion of the newly generated problem of work and income distribution, THE MECHANIZATION OF WORK is a penetrating analysis of this major force in our lives.

1982, 119 pages, 114 illustrations
hardbound: ISBN 0-7167-1438-8 $19.95
paper: ISBN 0-7167-1439-6 $9.95

How to Order: Simply complete this coupon and mail it together with payment or credit card information to the publisher at the address below.

THE MECHANIZATION OF WORK. hardbound 1438-8 _____ copies @ $19.95 _____

paper 1439-6 _____ copies @ $ 9.95 _____

Postage & handling charge 1.50

California, Utah & New York residents add appropriate sales tax _____

☐ I enclose a check or money order payable to W. H. Freeman and Company. Total $ _____

☐ Please charge my ☐ MasterCard ☐ VISA

Credit Card # _____ Expires _____ Signature _____
(all credit card orders must be signed)

Name _____ Address _____

City/State _____ Zip _____

Also available at fine bookstores.

W. H. Freeman and Company, 660 Market Street
San Francisco, CA 94104 (SA3)

Hot Springs on the Ocean Floor

A sign of the volcanic activity that mints new oceanic crust, they nourish strange forms of life and lay down great deposits of ore. They also help to explain the chemical content of the sea

by John M. Edmond and Karen Von Damm

Hot springs are common at the bottom of the ocean along the submarine ridges where large areas of the surface of the earth are being driven apart and new oceanic crust is rising into place. One might have made that prediction soon after the formulation of the theory of plate tectonics, which explains the significance of the ridges. Indeed, the prediction was made, by J. W. Elder of the University of Manchester. Now, two decades later, it is being confirmed. Cameras, sensors and research submarines operating at ocean depths approaching three kilometers are finding large numbers of hot springs, and the water they vent into the sea is proving to have remarkable consequences.

First, the chemical content of the vent water, which originates in a complex set of reactions between seawater and the hot, newly minted crustal rock, is proving to nourish a chain of living organisms extending from bacteria to clams and giant tube-dwelling worms. Second, the hydrothermal reactions are proving to be the main source of the metal-rich sediments and nodules that carpet the floor of the ocean. The exploitation of these deposits is now under worldwide debate. Other ore deposits, which lie on the continents and have been exploited since ancient times, are proving to have been emplaced there by plate-tectonic motions after hydrothermal activity had manufactured them under the sea. Third, the chemical content of the vent water is proving to be an important determinant of the chemical composition of the oceans themselves.

In retrospect the place of hot springs in the theory of plate tectonics is quite straightforward. In the theory the surface of the earth is viewed as a set of large, rigid plates that are in constant motion with respect to one another. The plates themselves are mostly the crust underlying the oceans, with great rafts of lighter material, the continents, embedded in them. At the margins where plates collide, old crust is being consumed. Such margins are trenches where plates plunge into the plastic interior of the earth's mantle. At the margins where plates diverge, new crust is being created. These latter margins are almost entirely under the oceans. They consist of linear features on the order of 100 kilometers long, offset in a zigzag pattern by faults a few kilometers long.

The Midocean Ridges

Magma (molten rock produced by the partial melting of the mantle at depths of no more than a few hundred kilometers) rises into the linear features. Its temperature as it arrives is some 1,200 degrees Celsius. It cools and solidifies, forming new oceanic crust; then, as the cooling continues, the crust is slowly rifted apart, making way for newer intrusions. The process finds expression in the height of the ocean floor. Hot material expands; hence the intrusion zones are high. They are the midocean ridges, which top off at depths between 2.5 and three kilometers in the sea. Cooling material contracts; hence the ocean floor subsides as it moves away from the ridges. As a result most of the ocean floor (and therefore most of the surface of the earth) obeys a gratifyingly simple relation: its altitude decreases with the square root of its age. In the Atlantic the part of the sea floor that is 100 million years old has moved at a rate of about one centimeter per year to its present position about 1,000 kilometers from the mid-Atlantic ridge, where it formed. As it has cooled it has subsided about two kilometers, giving the ocean there a depth of five kilometers.

One readily imagines that since the tectonic spreading centers at the midocean ridges are places of volcanic activity, they, like Yellowstone National Park and other areas of volcanic activity on the continents, ought to have hot springs, the sign of hydrothermal activity. Still, one wants to have evidence. The evidence soon came to hand.

In 1966 Kurt G. T. Boström and Melvin N. A. Peterson of the Scripps Institution of Oceanography analyzed sediment samples that had been recovered at sites on each side of the axes of midocean ridges. In this way they found that the newly identified oceanic spreading centers were covered by sediments rich in oxides of iron, manganese and other metals. They proposed that the sediments resulted from hydrothermal reactions of seawater with hot, young oceanic crust. Subsequent work by the Deep Sea Drilling Project, administered by the Scripps Institution, has shown that metalliferous sediment is ubiquitous at the base of the ocean's sediment column, immediately above the volcanic rock of the oceanic crust.

Meanwhile the theory of plate tectonics was pushing the rock of the oceanic crust into the foreground of geologists' attention. The oceanic crust is basalt, a black volcanic rock. Chemically it is an aluminum silicate ($AlSiO_3$) free of quartz (SiO_2) but rich in iron and manganese. (In contrast, granite and rhyolite, the volcanic rocks typical of the continents, are whitish aluminum silicates poor in iron and manganese. In essence they are the slag that has risen to the top of an immense smelting furnace, the earth.)

In the 1960's dredging campaigns

BLACK SMOKER at ■ midocean ridge is the most dramatic expression of hot springs and hydrothermal activity at the bottom of the ocean. The smoker shown here was photographed from the research submarine *Alvin* at ■ depth of 2.6 kilometers in the Pacific just south of the entrance to the Gulf of California. It lies on the ridge of the East Pacific Rise. Its diameter is about 1.5 meters; the temperature of the water it vents is 350 degrees Celsius. The water has permeated the oceanic crust, reacted with hot basaltic rock in ■ zone of volcanic activity under the East Pacific Rise and returned to the ocean floor ■■ ■ hydrothermal solution: ■■ acidic liquid rich in metals. On its return the solution is cooled by seawater and blackened by ■ precipitate of iron sulfide particles. The smoker chimney forms from minerals that also precipitate.

were launched to sample outcrops close to ridge axes. The Deep Sea Drilling Project recovered short cores of rock from the bottom of drill holes. The analysis of these samples showed that in general they had been subjected to at least one episode of chemical reaction with seawater after their emplacement as oceanic crust. For example, cracks had developed in the rocks as they cooled to a temperature of less than, say, 500 degrees C., and evidently seawater had permeated the cracks, causing chemical reactions and filling the cracks with precipitated, or secondary, minerals.

This was confirmed by isotopic data. In any chemical reaction involving oxygen the isotopes, or nuclear species, of oxygen are fractionated: they are distributed somewhat inhomogeneously among the reaction products. The fundamental reason is that the vibrational energy of a chemical bond depends on the mass of the atoms bonded together. The degree of fractionation is characteristic of the particular reaction; it also depends on the temperature. The analysis of secondary minerals in the basalt samples from the oceanic spreading centers showed, then, that their composition of oxygen isotopes differed from that of the basalt. In fact, it showed that the minerals had formed by reactions of the basalt with seawater at temperatures as high as 350 degrees C.

Ophiolites

Further evidence of hydrothermal activity at oceanic spreading centers was gathered, surprisingly enough, on the continents. Great slabs of basaltic rock found on top of continental crust on the island of Cyprus, on the northeastern Mediterranean littoral, in Oman on the Arabian Sea, in California, in Tibet and elsewhere were identified as segments of oceanic crust that had been driven onto the continents by plate collisions. Called ophiolites, these formations provided an opportunity to study the mechanisms of formation of the sea floor in detail. In ophiolite terrains one could readily sample and map what had once been oceanic crust, often over hundreds of kilometers horizontally and as much as 10 kilometers vertically. Again the signs of large-scale, high-temperature hydrothermal activity were impressive. In the great ophiolite of Oman, for example, it became plain that seawater had once penetrated to depths in excess of five kilometers and reacted pervasively with the rock at temperatures of about 400 degrees C.

In most ophiolites the basalts are covered by a metalliferous sediment, called umber, that is several meters thick and is identical with the ones Boström and Peterson found on the present-day ocean floor. Moreover, lenticular (lens-shaped) ore bodies containing millions of tons of iron sulfide (FeS_2, known as pyrite or fool's gold) are common in ophiolites. They occupy depressions in the surface of the basalt and are underlain by "pipes" of quartz and ore minerals. Clearly the pipes were once conduits

GIANT CLAMS AT A HOT-SPRING FIELD on the ocean floor just south of the Gulf of California cluster between pillow-shaped mounds of basalt at fissures from which hydrothermal solutions are venting at a temperature of about 17 degrees C. The solutions are

for ascending hydrothermal solutions from which the quartz, the ore minerals and the lenticular bodies precipitated.

In an effort to directly determine the heat balance of the oceanic crust as it evolves, measurements of the conductive loss of heat from the sea floor have been gathered from all parts of the oceans. The conductive loss represents the flow of heat through a motionless medium, in this case the sediment covering the sea floor. In general the conductive loss turns out to decrease toward the midocean ridges. Yet the crust at the ridges must be hotter than the older crust farther away. Hence much of the heat brought to the ridges by ascending magma is being removed by another mechanism, namely convection, the transport of heat by a fluid, in this case seawater. Indeed, Clive R. B. Lister of the University of Washington has calculated that the thermal processes at the ridge axis must be dominated by convection.

The global integration of the anomaly between the measured heat loss and the amount one would expect if the loss were entirely conductive gives a value of about 5×10^{19} calories per year, or about a tenth of the total heat flux from the interior of the earth. This value must be attributed to convection. Suppose the operating temperature of the convective cells is 350 degrees C., in agreement with the isotopic data. Then a volume of seawater equal to the volume of the oceans (1.37×10^{21} liters) must circulate through the ridge axes, and therefore through a temperature regime of 350 degrees C., every eight million years or so. The rate is about half of 1 percent of the rate at which the oceans receive water from the rivers of the earth.

rich in hydrogen sulfide, which nourishes certain bacteria. The bacteria nourish the clams.

Helium Isotopes

Perhaps the strongest evidence that hydrothermal activity is important in the oceans today comes from measurements of the concentration of helium in the oceanic water column. Helium has two isotopes: helium 3 and helium 4. At the time the earth was formed the two isotopes had a certain ratio of abundance with respect to each other. Helium 4, however, is produced inside the earth by the radioactive breakdown of the long-lived isotopes of uranium and thorium. Moreover, both helium 3 and helium 4 are escaping into interplanetary space from the upper atmosphere of the earth, and at the same time both of them are being injected into the atmosphere by the wind of particles that arrives at the earth from the sun.

The ratios of helium 3 with respect to helium 4 from place to place on the earth represent, therefore, the outcome of several processes. Brian Clarke of McMaster University has developed a sensitive mass spectrometer that accurately measures the isotopic composition of helium, and he has employed the device to analyze the helium dissolved in water that trickles from aquifers into wells. The concentration of helium 3 in the water is assumed to have undergone no change since the time the water first entered the aquifer; thus it preserves a record of the equilibrium solubility of helium: the amount that naturally passes into the water from the atmosphere. Any anomaly in the concentration of helium 4 with respect to this equilibrium value can be taken to result from subsurface radioactive decay. In short, the spectrometer is an aid to uranium prospecting.

With that much established, Clarke, in collaboration with Harmon Craig of Scripps, attempted to estimate the leakage of helium 4 from the mantle into the ocean. As before, they assumed that helium 3 would serve as an index. What they found instead was remarkable. At the surface of the Pacific the helium dissolved in the sea was in fact in equilibrium with the helium in the atmosphere. At greater depths the amount of helium exceeded the equilibrium value, and when the equilibrium value was subtracted from the measured concentration, the amount of helium 3 in the excess was anomalously large. The anomaly attained its maximum at 2,600 meters, the average depth of the Pacific ridge axis. There the ratio of helium 3 to helium 4 in the excess helium was about eight times the atmospheric value. Evidently helium 3 is retained in the mantle and is being released into the ocean at the ridge axis. Does the helium simply escape from gas bubbles near the surface of cooling magma or is it liberated in the course of hydrothermal activity that pervasively reworks the rock? The scale of the anomaly strongly indicated the latter.

The Composition of Seawater

Precisely how does seawater react with basalt at sea-floor pressures of almost two tons per square inch and at temperatures of several hundred degrees C.? The first answers emerged in the mid-1970's when James L. Bischoff and Frank W. Dickson reproduced those conditions in their laboratory at Stanford University. The results were dramatic. Magnesium ions in the seawater combined with silicate in the basalt to form a hydroxysilicate, $Mg(OH)SiO_3$, which is insoluble in water. The hydroxyl (OH) groups were contributed by the water, leaving hydrogen ions (H^+) behind; thus the seawater solution became highly acidic. Hydrogen ions then took the place of calcium and potassium at sites in the basalt's crystal lattice, and so the lattice recrystallized.

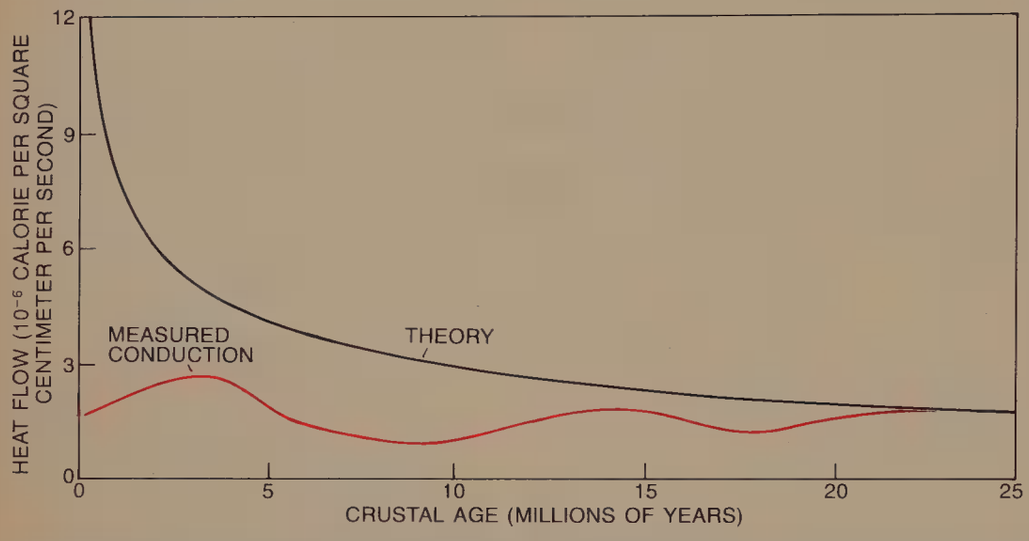
In essence the reactions were an exchange of hydrogen ions for other cations, or positive ions, notably those of calcium and potassium. The liberated calcium could now react with sulfate (SO_4) in the seawater to form calcium sulfate ($CaSO_4$), which precipitates as the mineral anhydrite. Sulfate could also react with iron in the basalt to form iron sulfide, that is, deposits of pyrite. The result of these two reactions was the complete removal of sulfate from the seawater. Finally, the degradation of the basalt and the acidity of the seawater solution allowed the release of silica (SiO_2 dissolved in water).

Might the chemical composition of the oceans come as much from these hydrothermal reactions as it does from the products of weathering on the continents? The possibility was particularly exciting because in the mid-1970's the understanding of the processes that control the composition of seawater were at an impasse. In the 19th century, before the boom in organic chemistry that followed the discovery of synthet-

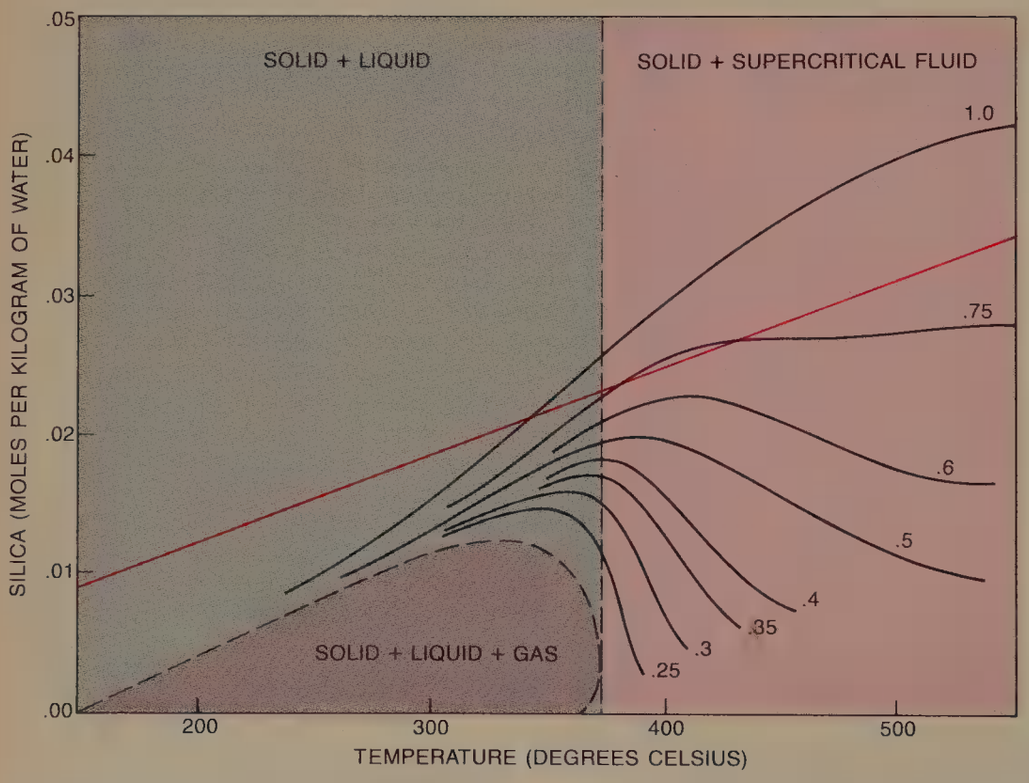
ic dyes, many prominent chemists had undertaken analyses of inorganic natural substances.

One important motivation was the search for unknown chemical elements to fill the gaps in the periodic table of the elements, which was then being developed. The chief strategy was the analysis of ore minerals. R. W. Bunsen, for example, analyzed hundreds of ores. He became interested in magmatic processes and traveled to Iceland to witness volcanic eruptions. Other investigators directed their attention to the study of natural waters. Thus J. H. van't Hoff laid the foundation of experimental geochemistry by studying the sequence of precipitates that form as seawater evaporates.

From all these efforts it was plain by the end of the century that seawater could not be produced by the partial evaporation of river water. At the end of that route lie only closed-basin lakes such as the Dead Sea and Great Salt Lake, which are highly alkaline compared with the oceans. Then came a long hiatus. It lasted until the late 1950's, when Lars Gunnar Sillén of the Royal Institute of Technology in Stockholm undertook to clarify matters. Sillén asked: What controls the *p*H, or hydrogen-ion concentration, of the oceans? Why is it consistently 7.5 to 8, or quite close to acid-alkaline neutrality? Subsequently Frederick T. Mackenzie and Robert M. Garrels of Northwestern University translated Sillén's rather abstract answer into geological terms.



FLOW OF HEAT from the ocean floor was an early sign that hydrothermal activity is important at midocean ridges. At each ridge, plates of oceanic crust diverge and new crust rises into place; hence the ridges are notably hot. Yet the measured conduction of heat through the crust at three Atlantic ridges (*color*) proved to be markedly less than the amount to be expected if the flow were entirely by conduction (*black*). The difference must be carried by conduction, that is, water passing through the ocean floor. The horizontal axis of the chart shows the age of the oceanic crust, a correlate of its distance from the midocean ridge where it was emplaced.

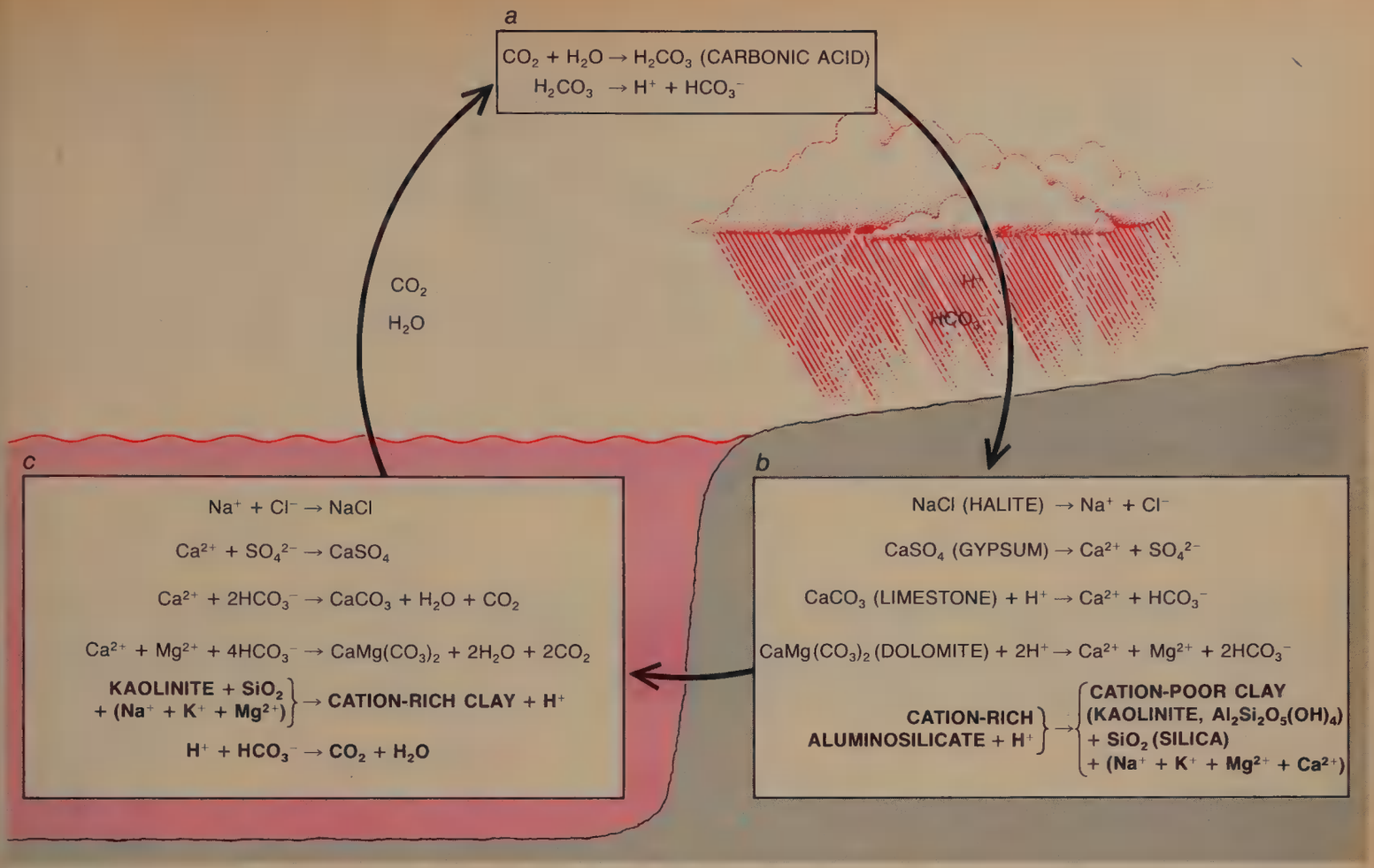


SOLUBILITY OF QUARTZ in vent water sampled in 1977 at ocean-floor hot-spring fields near the Galápagos Islands suggested even before black smokers were discovered (two years later) that pure hydrothermal solutions rising toward the ocean floor might turn out to have a temperature of about 350 degrees C. The Galápagos samples were no hotter than 19 degrees; still, the concentration of silica (dissolved quartz) was greatest in the hottest samples. The extrapolation of the trend (*color*) intersected curves representing water saturated with quartz (*black*) at points well above 300 degrees. (Pressures are labeled in kilobars.) Evidently the Galápagos samples represented hydrothermal solutions diluted with ordinary seawater. The chart shows the regimes in which a quartz-water mixture is stable as (1) solid plus liquid plus gas, (2) solid plus liquid and (3) solid plus supercritical fluid (a phase that is neither liquid nor gas).

Consider the weathering of volcanic rocks on the continents. The rain beating down on the rocks is an acid solution because carbon dioxide from the atmosphere dissolves in water to make carbonic acid (H_2CO_3). For their part the rocks are an aluminum silicate lattice in which cations are embedded. The acid reacts with the lattice. Specifically, the acid gives up hydrogen ions, leaving bicarbonate (HCO_3^-) behind; thus the rainwater becomes alkaline. The hydrogen ions take the place of cations in the lattice, freeing them into solution. The lattice is therefore disrupted; it is transformed into highly disordered clay minerals such as kaolinite ($Al_2Si_2O_5(OH)_4$), which are easily eroded.

Estimates of the global rate of these reactions can be derived from the rate at which their products are carried toward the oceans by the rivers. The result requires that an amount of carbon dioxide equal to its abundance in the atmosphere be consumed in about 4,000 years. Clearly there must be a reaction that converts bicarbonate back into carbon dioxide. Mackenzie and Garrels noted that the ratio of cations to hydrogen ions is about 1.2×10^3 in average continental waters. It is 6×10^7, four orders of magnitude greater, in the oceans. Thus when aluminum silicates rich in hydrogen ions are transported into the oceans (where hydrogen ions are scarce and cations are plentiful), the reactions that made them should reverse. Clays rich in cations should form; hydrogen ions should be released and then should combine with bicarbonate to make water and carbon dioxide.

The implications were profound. The gross composition of the continental crust has been constant for at least the past 2.5 billion years; hence the composition of the oceans would also have been unchanging over that vast period. Studies were undertaken in an attempt to establish that the crucial reactions re-



CHEMICAL BALANCE between the atmosphere, the continents and the oceans was hypothesized in the 1950's by Lars Gunnar Sillén of the Royal Institute of Technology in Stockholm and worked out in detail by Frederick T. Mackenzie and Robert M. Garrels of Northwestern University. In the atmosphere (*a*) carbon dioxide is dissolved in water vapor as carbonic acid. The water rains or snows on the continents (*b*), where hydrogen ions from the acid solution leach cations (positive ions) such as calcium, magnesium, sodium and potassium out of the continental rocks. Rivers carry the results of the continental weathering to the oceans (*c*). In the oceans, which are poor in hydrogen ions, the reactions reverse, creating sedimentary rock rich in cations and restoring carbon dioxide to the atmosphere.

constituting clay minerals were in fact happening in the oceans. The findings were disappointing. Even the most optimistic interpretation of the data could not approach the scale necessary to close the balance between the land and the sea. Here, then, was the impasse mentioned above. The preliminary calculations of the chemical fluxes to be expected from ridge-axis hydrothermal activity pointed to a way out.

First Explorations

Of course, being hopeful that hot springs are common at the oceanic spreading centers and actually finding them under at least 2.5 kilometers of water are very different things. When the first attempts were made in the early 1970's, knowledge of the sea floor and the capacity to explore it were equally rudimentary. The sonar devices then in service were adequate for flat terrain. At the rugged topography of the ridge axes, however, they recorded a jumble of echoes. As a result the equipment towed close to the bottom in efforts at exploration was in effect pulled through elaborate obstacle courses in the dark. Instruments were damaged and lost at an alarming rate. Nevertheless, the most sophisticated device, Deep Tow, operated by the Scripps Institution, made significant findings.

Deep Tow, a vehicle towed at the end of a telemetry cable from a ship on the surface, carried television cameras, sonar, pressure sensors and probes to measure seawater temperature and electrical conductivity (a correlate of salt content). Over several years it detected temperature anomalies at a number of places on the spreading centers in the eastern tropical Pacific. In one case Deep Tow carried water samplers, and Ray F. Weiss of Scripps managed to catch a sample from the midst of an anomaly. The temperature of the water differed by less than .1 degree C. from the temperature of the ambient water: two degrees. Still, the chemical data, including the concentration of helium 3, showed without doubt that the water was of hydrothermal origin. Presumably Deep Tow had "flown" into a plume 15 or 20 meters above a submarine hot spring.

In the mid-1970's changes began to come rapidly. For one thing, the U.S. Navy made available to researchers the techniques it had developed for mapping the ocean floor. Thus high-precision deep-sea navigation systems came into routine service. Such systems depend on the measurement of the time that elapses between the broadcast of an acoustic pulse (an "interrogation pulse") from an undersea vehicle and the reception of acoustic "replies" from an array of transponders moored to the bottom. If the relative positions of the transponders have been determined by sonar from a ship at the surface, the undersea vehicle can be navigated quite readily. Indeed, positional accuracies of a few meters are achievable.

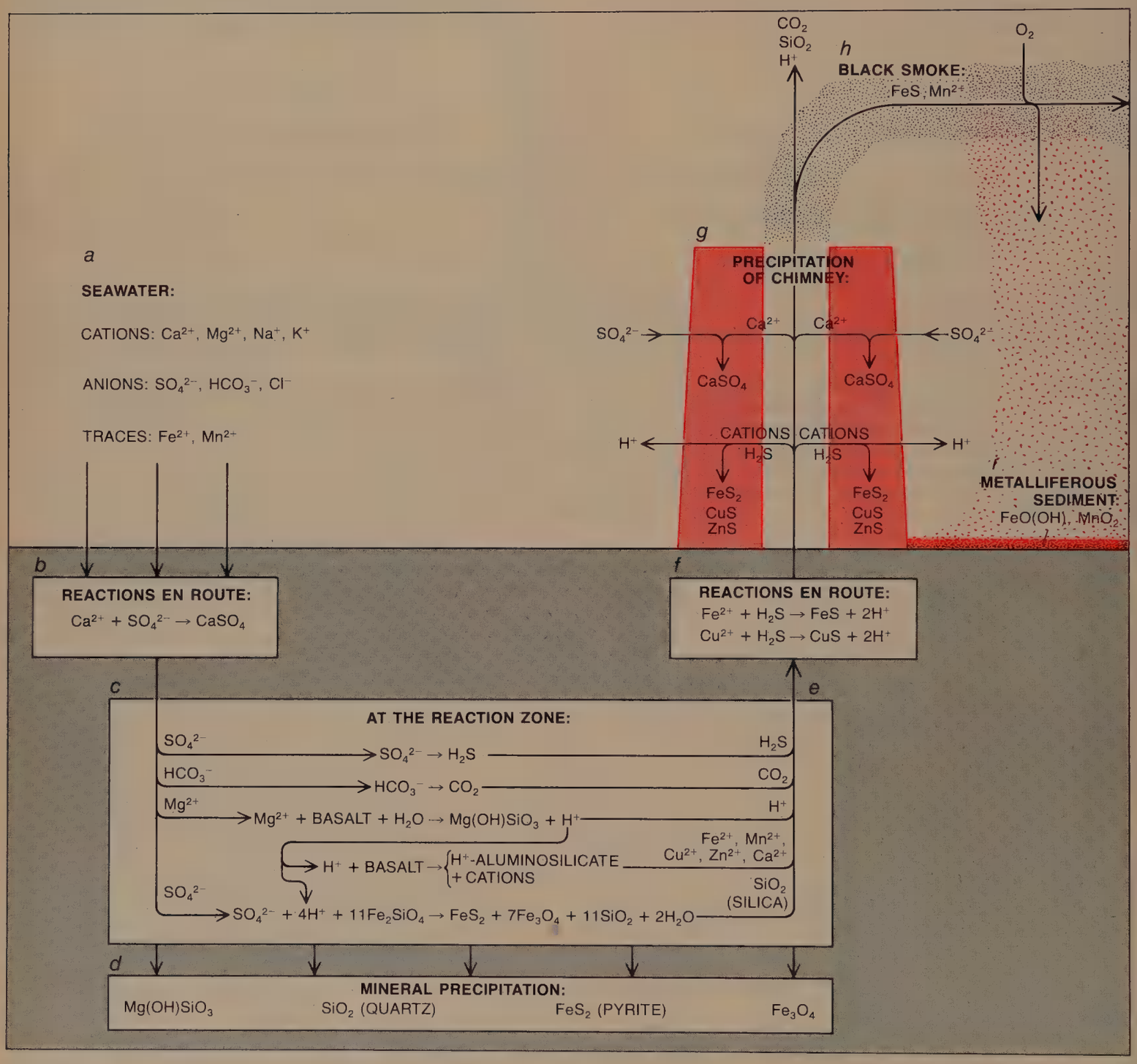
Second, a large-field survey camera was developed. The available cameras had been flimsy, lightweight devices towed far behind a ship, so that they were difficult to control. What was needed was a massive camera vehicle. Angus, developed at the Woods Hole Oceanographic Institution, was the result. It is a 1.5 ton "gorilla cage" in which are mounted color cameras, stroboscopic lamps, power supplies, sonar and an acoustic navigation transponder. In its current configuration the vehicle is towed from a ship at a rate of about four kilometers per hour at elevations of about 20 meters above the sea floor. Because of its mass, it is never more than 75 meters from being directly under the ship; hence the ship's sonar read-

ings can help to keep the vehicle from harm. Normally a tow of Angus lasts for 18 hours. A color photograph is made every 10 seconds. The vehicle is brought up, the film is developed and intriguing features are noted. Each photograph bears a record of when it was made; from this datum the feature can be located exactly.

The third crucial technology is the research submarine. The submarines that can work at ridge-axis depths are small. The largest of them, *Alvin,* operated by Woods Hole, weighs 16.5 tons and carries two investigators and a pilot in a titanium pressure-resistant sphere two meters in diameter. It is severely limited in power. Thus it is slow: its maximum speed over the bottom is about four kilometers per hour. Moreover, its lights can illuminate the water to a distance of only about 15 meters. It is an inefficient tool for exploration. Its proper role, however, is in visiting sites already identified (say in photographs made by Angus) as deserving study. *Alvin* is directed to such targets by means of the same navigation transponders that guide Angus. Routinely one can land *Alvin* on the sea floor within a few tens of meters of the designated target.

A Fabulous Scene

This elaborate and costly three-part system was first deployed in the spring of 1977 at a spreading ridge in the Pacific 280 kilometers northeast of the Galápagos Islands. The research vessel



HYDROTHERMAL REACTIONS at the midocean ridges regenerate carbon dioxide to a far greater extent than the low-temperature oceanic process Sillén envisioned. Seawater (*a*) percolates into the oceanic crust. As it descends some of its content of ions may precipitate (*b*) as calcium sulfate. The remaining ions reach ■ zone several kilometers under the ocean floor. There they react with hot basaltic rock (*c*) to produce mineral precipitates (*d*) and ■ hot, acidic, metal-rich hydrothermal solution (*e*) that rises back toward the ocean floor. The rising solution may encounter cold seawater and cool, thus precipitating metal sulfides (*f*). Then, at the ocean floor, the solution mixes with sulfate from seawater to precipitate chimneys consisting of metal sulfides and calcium sulfate. Above the chimneys the iron still in solution forms black iron sulfide smoke (*h*). Manganese stays in solution. Eventually the iron and the manganese are oxidized and rain onto the ocean floor as ■ metalliferous sediment that is ubiquitous atop the oceanic crust (*i*). Meanwhile the carbon dioxide emerging from the vents is mixing into the ocean. Eventually it reaches the surface, enters the atmosphere and closes the carbon dioxide cycle.

The world is full of guarantees, no two alike. As a rule, the more words they contain, the more their protection is limited. The Lands' End guarantee has always been an unconditional one. It reads:

> *"If you are not completely satisfied with any item you buy from us, at any time during your use of it, return it and we will refund your full purchase price."*

We mean every word of it. Whatever. Whenever. Always. But to make sure this is perfectly clear, we've decided to simplify it further.

GUARANTEED. PERIOD.

LANDS' END
DIRECT MERCHANTS

of fine wool and cotton sweaters, Oxford button-down shirts, traditional dress clothing, snow wear, deck wear, original Lands' End soft luggage and a multitude of other quality goods from around the world.

☐ **Please send free catalog.**

Lands' End Dept. Q-16
Dodgeville, WI 53533

Name _____

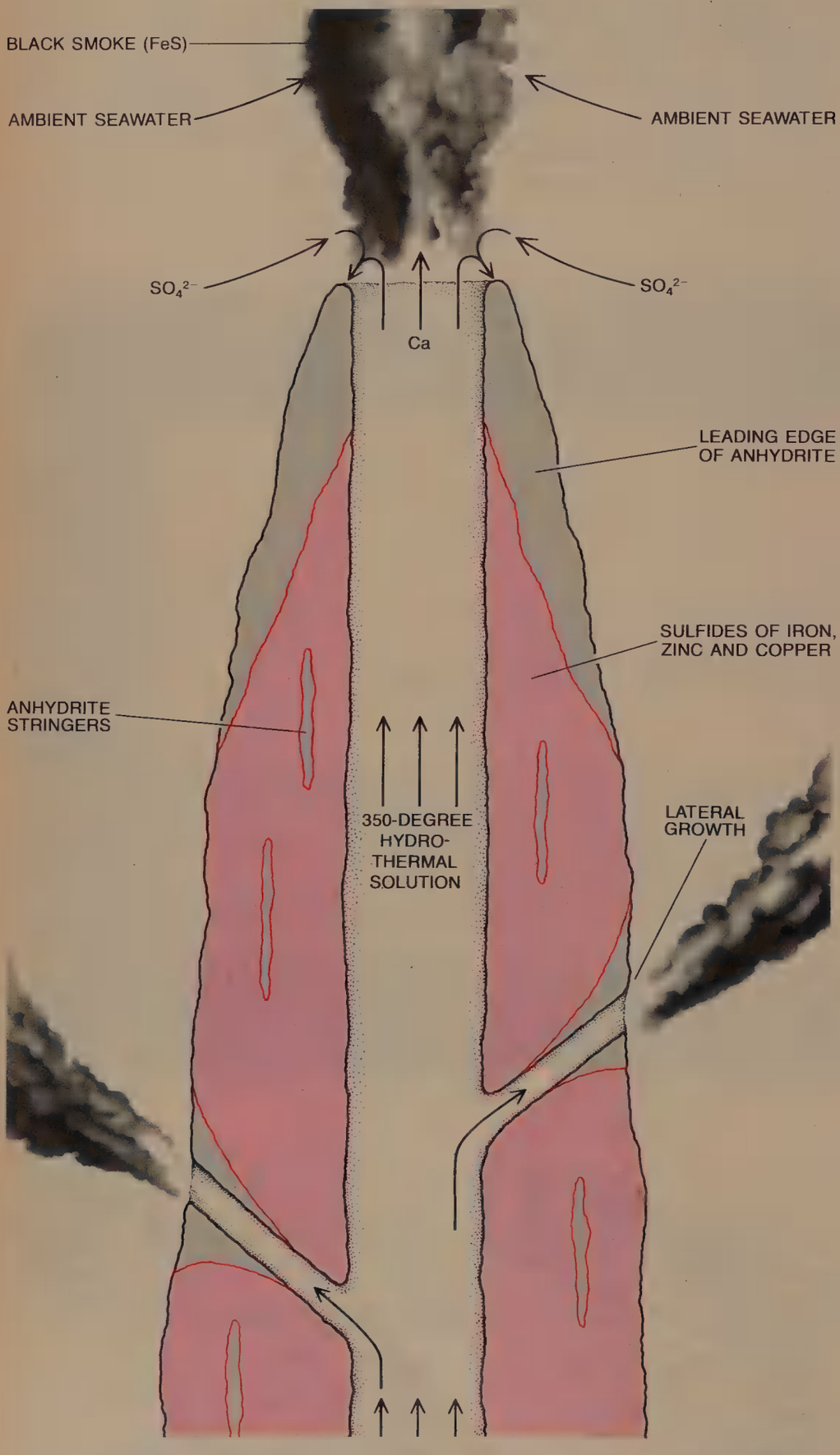
Address _____

City _____

State _____ Zip _____

Or call Toll-free
800-356-4444
(Except Wisconsin, call 608-935-2788)

Knorr, out of Woods Hole, arrived first. It laid down a network of transponders and surveyed their positions. Then Angus began its runs, under the direction of Robert D. Ballard of Woods Hole. By the time *Alvin* arrived a few days later on its tiny mother ship *Lulu,* several targets had been identified. At each one the Angus photographs showed a few large white clams on a black basaltic field. *Alvin* was directed to one of the targets. The investigators aboard were John B. Corliss of Oregon State University and one of us (Edmond).



BLACK SMOKER ACCRETES from calcium sulfate, which precipitates out of a hydrothermal solution as a "leading edge" of the whitish mineral anhydrite. As the leading edge advances it becomes exposed to hot, undiluted solution. As it redissolves, however, it acts as a template for metal sulfides. Anhydrite is preserved in places where sulfides engulf it rapidly.

Alvin sinks passively in the ocean's water column at a rate of 30 to 35 meters per minute. Thus it settled to a depth of 2,500 meters in one and a half hours. At that point, 100 meters from the bottom, it jettisoned a set of weights and gained neutral trim. We moved downward until we hovered above the sea floor, which proved to be a gentle slope. For about half an hour we cruised about, trying to find the target. Each of us had a Plexiglas porthole to look through. Then we stopped to collect some rocks. As our pilot employed the submarine's mechanical arm to wrestle with a basalt pillow (a volcanic deposit that results from the slow extrusion of lava on the sea floor when the surface of the lava chills, hardens, cracks and heals so that the lava inflates itself into the shape of a pillow) a couple of large purple sea anemones engaged our attention. Only when our gaze finally wandered away from them did we realize that the water within the range of our lights was shimmering, like the air above a hot pavement. The hastily measured temperature of the water was five degrees above the ambient water temperature (2.05 degrees C.). With all thoughts of rocks forgotten, we captured a sample of the water and then continued on our course upslope. Here we came on a fabulous scene.

The typical basaltic terrain at the ridge axis is bleak indeed. Monotonous fields of brown pillows are cut by faults and fissures. One must examine several square meters to find a single organism. Yet here was an oasis. Reefs of mussels and fields of giant clams were bathed in the shimmering water, along with crabs, anemones and large pink fish. The remaining five hours of "bottom time" passed in something close to frenzy. We recorded the temperature, conductivity, *p*H and oxygen content of the water; we made photographs; we sampled the water; we made sure that a representative selection of organisms was collected, all under the growing pressure of steadily decreasing voltages for our equipment.

Fortunately the equipment worked flawlessly. It soon became apparent that we had come on a hot-spring field. Warm water streamed from every orifice and crack in the sea floor over a circular area about 100 meters in diameter. The temperature of the water was highly variable, but the maximum was about 17 degrees C. The organisms were quite selective. They choked the warmest vents. In some cases mussel reefs actually channeled the flow of water, forming conduits themselves.

We worked until the "scientific power" ran out. Then a second set of weights was jettisoned and *Alvin* left the bottom. The equipment was shut down. Soon we began to get cold. After eight hours of crouching in the two-meter sphere, with almost every movement requiring the

Sale! Save $500 on Radio Shack's Two-Disk TRS-80® Model III Computer

Only 1795.00

Reg. 2295.00

48K 2-Disk with RS-232C
(Cat. No. 26-1066)

TRS-80—The Best Supported Microcomputer in the World. Save $500 on the professional personal computer sold and serviced through over 8700 locations worldwide. Radio Shack's Model III can streamline your work or your staff's without writing programs, because we offer more ready-to-run software than any other microcomputer manufacturer!

Save Time and Effort. Answer "what if?" questions with VisiCalc®, an electronic spreadsheet. Great for formula analysis, statistical research, design analysis, pricing analysis, budget planning and much more. Create easy-to-understand charts and graphs with a plotter or printer and our graphics analysis program. Our "Manager Series" helps you to plan, organize, schedule and record your business and personal affairs. Keep track of project files, personnel records, technical data and more with our electronic filing program, Profile® III Plus. Produce correction-free memos, articles and reports with SuperSCRIPSIT™. This powerful word processing program supports proportional spacing, multiple column printing and super and subscripts to print mathematical formulas.

The "Extras" Are Built-In. The TRS-80 Model III is complete and self-contained. You get a 48K internal memory, a 12" high-resolution monitor, which displays 16 lines of 64 upper and lower case characters, a 65-key typewriter-style keyboard with numeric datapad, and two built-in double-density 5¼" disk drives that provide 368K of on-line program and data storage. The RS-232C Serial Communications Interface makes it easy to access corporate or national information networks. A parallel printer interface is standard, too. And Model III is fully expandable to meet your growing needs. You can add high-resolution (640 x 240) screen graphics, a 5-megabyte hard disk drive, digitizer and more. You can use optional programming languages like COBOL and FORTRAN, too.

Hurry—Sale Ends April 25, 1983. Get this TRS-80 today at your nearest Radio Shack Computer Center, store or participating dealer. Our cassette-based Model III is on sale, too. It's a great starter system for only $799 (26-1062, regularly $999). Ask about our leasing, service and training plans—and pick up a free copy of our new 1983 TRS-80 Computer Catalog!

Radio Shack®

The biggest name in little computers®

A DIVISION OF TANDY CORPORATION

Retail prices may vary at individual stores and dealers. Special order may be required. VisiCalc is a registered trademark of VisiCorp.

cooperation of one's equally cramped colleague, we spent the 90-minute ascent trying to keep warm, too tired even to watch the bioluminescent organisms disturbed by the submarine's passage. The approach to the surface, as always, was like a rebirth. At a depth of about 200 meters a faint green glow began to pervade the water. Soon all was bright. The pilot vented the ballast tanks, surrounding *Alvin* with a column of bubbles. There was a soft bump as we surfaced; then the lazy roll in the swell. Shortly divers were around the submarine attaching recovery lines. *Alvin* was maneuvered onto its recovery cradle. Then came the agony of standing up and clambering out. By the time we had staggered onto *Lulu,* everyone shared our excitement. *Alvin's* sample basket was loaded with clams and mussels. Crabs, which had got inside the fiberglass skin of the submarine, began to fall out on deck.

Living on Hydrogen Sulfide

Now the real effort began. The water samplers were disassembled. The water samples were transferred to the *Knorr* and analyzed through the night. The Angus photographs were scrutinized for more targets. The water samplers were cleaned and reassembled. At 6:00 A.M. the next day *Alvin* was being readied; at 8:45 it descended again. It made a total of 15 dives.

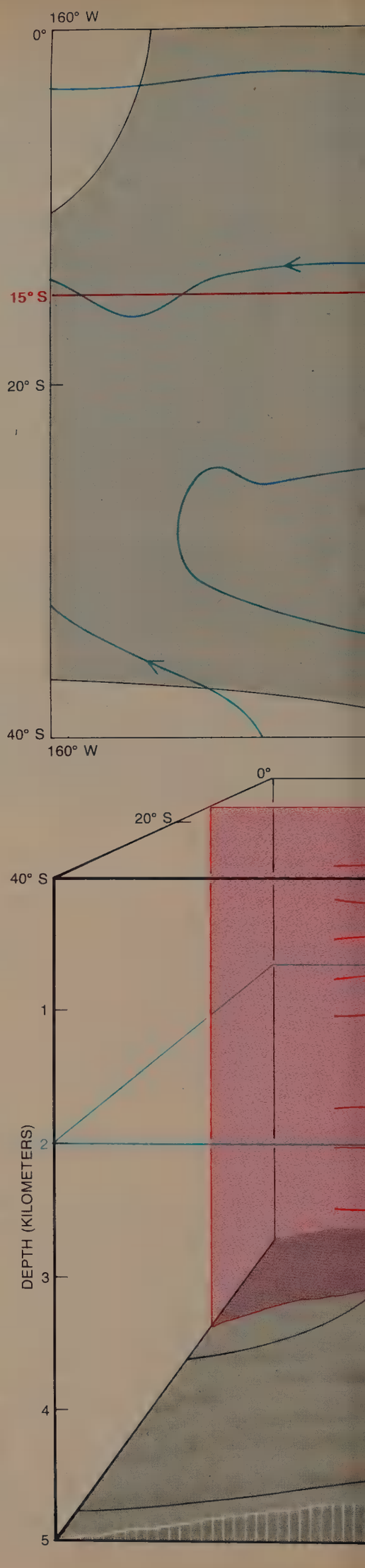
As the analytical work progressed it became clear that all the hot-spring waters had a high concentration of hydrogen sulfide (H_2S). Herein lay the explanation of the ocean-floor oases. Bacteria that derive their energy from the oxidation of hydrogen sulfide are common in many ecological systems. At the oases, however, they had to be the primary producers at the base of the ecological pyramid. Fundamentally the energy driving the newly discovered ecosystem flows not from the sun, as in photosynthesis, the source of energy for all other ecosystems, but from the radioactive breakdown of long-lived isotopes of uranium, thorium and potassium inside the earth. The breakdown releases heat; the heat produces magma, which becomes new oceanic crust. Seawater percolates into the crust and reacts with the crustal rock at high temperature and pressure.

Two of the reactions are crucial for the oases. Sulfate from the seawater reacts with iron in the rock to produce hydrogen sulfide and iron oxides; moreover, sulfide minerals in the rock are dissolved. Thus the hot water rising back toward the ocean floor is heavily charged with sulfide. At vents on the sea floor the hydrogen sulfide is absorbed by bacteria that also take up oxygen of photosynthetic origin dissolved in the ambient water. The sulfide and the oxygen combine, so that the sulfide is turned back into sulfate. The reaction releases the energy that drives the bacteria's metabolism. The bacteria in turn nourish other species. The diversity of the species was an immediate indication that hot springs must be common on the ridge axes of the world. How else could a unique and highly evolved fauna be developed and sustained?

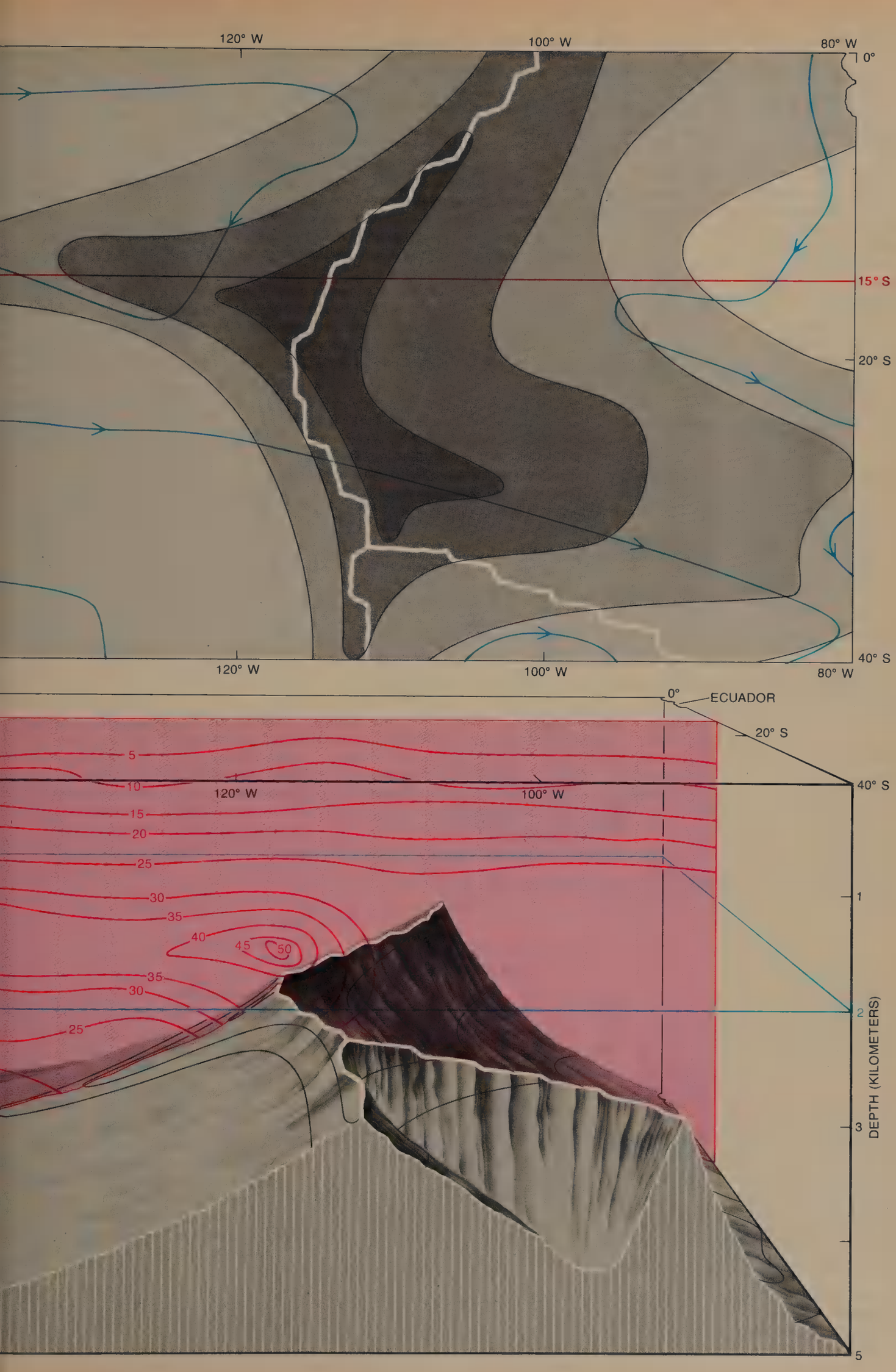
Our own laboratory at the Massachusetts Institute of Technology had been assigned to analyze in detail the water *Alvin* had sampled at the vents. Plainly the chemistry of the water was the result of a complex set of reactions between seawater and basalt that proceed in an inaccessible reaction zone, perhaps several kilometers under the ocean floor. Hence the best way to go about unraveling the reactions would be to see how the water samples were enriched or depleted in the various chemical elements with respect to basalt and with respect to ordinary seawater. In the end we measured the concentration of 35 elements, probably the most ever measured in an effort of this type. As we progressed the chemical dynamics began to unfold.

The Vent Water's Properties

Magnesium and sulfate, which are constituents of ordinary seawater, showed a uniform decrease in concentration with increasing temperature



DISTRIBUTION OF METALLIFEROUS SEDIMENT on the floor of the South Pacific, as mapped by Kurt G. T. Boström of the Scripps Institution of Oceanography, has features that make it asymmetric with respect to the ridge axis from which the metals issue in a hydrothermal solution. The drawing at the top displays the South Pacific from above. At 15 degrees south latitude a long, narrow salient of sediment rich in oxides of iron and manganese projects westward from the axis; it is indicated in shades of gray. At 30 degrees south latitude a broader, shorter salient projects eastward. Blue lines mark ocean currents at a depth of 2,000 meters (about 600 meters above the ridge axis) as they have been deduced by Joseph L. Reid of the Scripps Institution. The currents have "blown" the oxides away from the axis; thus the currents account for the salients. The drawing at the bottom displays the South Pacific from the south. Red lines mark concentrations of the isotope helium 3 in a plume extending westward across the Pacific at 15 degrees south latitude. The pattern was measured by John E. Lupton and Harmon Craig, also of Scripps. Numbers indicate the concentration of helium 3 as a percent in excess of its normal saturation value in the sea. Evidently helium is freed from the oceanic crust by hydrothermal reactions. Because it is inert, it undergoes no change as it is vented at the axis; hence it labels the ocean water from which metalliferous sediment precipitates.

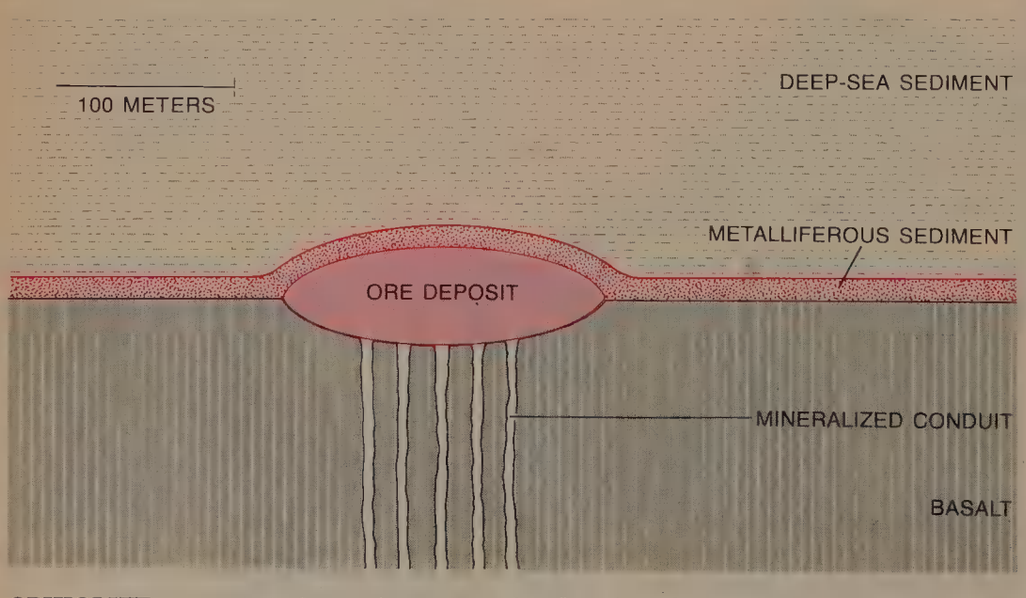

120° W
100° W
80° W
0°
15° S
20° S
40° S
ECUADOR
5
10
15
20
25
30
35
40
45
50
DEPTH (KILOMETERS)
1
2
3
5

(that is, the temperature of each water sample as *Alvin* had trapped it). Clearly each sample represented water that had issued from vents and then been diluted to some degree by the ambient seawater. The extrapolation of the concentration-temperature trends to a concentration of zero, on the hypothesis that pure vent water might prove to have no magnesium and no sulfate in it, yielded a temperature of about 350 degrees C. The extrapolation was admittedly extravagant: the hottest water we had encountered had a temperature of only 19 degrees. On the other hand, the concentration of silica (quartz dissolved in water) increased with increasing temperature. The extrapolation of the trend to its point of intersection with curves representing the solubility of quartz in water at various pressures and temperatures also yielded a temperature of about 350 degrees.

Iron, copper, nickel and cadmium, which combine with hydrogen sulfide in the vent water to form sulfides or oxides that precipitate out of water, all showed marked decreases with increasing temperature. Their trends extrapolated to zero at a range of temperatures from 30 to 35 degrees C. That range could be interpreted as the temperature of a reservoir of water being tapped by the hot springs. We surmised that the water ascending from a regime of high-temperature reactions entrained "groundwater" whose temperature and composition were close to those of the ambient seawater overlying the ridge axis. The high-temperature solution, or "end member," was cooled by this groundwater, and its acidity was diluted, causing the precipitation of the sulfide- and oxide-forming elements in the "plumbing" of the basaltic crust. The precipitation proceeded to such an extent that the sulfide- and oxide-forming elements were stripped even from the groundwater. They would then have been contributed to our samples only by the ambient water that diluted the samples. Evidence of the high-temperature reactions was preserved, however, in the magnesium, sulfate and silica data.

Helium 3 increased strongly with temperature, at a ratio of 2.2×10^{-17} moles (about 13 million atoms) for every calorie of heat. In the helium that remained when the equilibrium amount was subtracted from the measured concentration the ratio of helium 3 to helium 4 was eight times what it is in the atmosphere. In the wake of their discovery of the helium-3 anomaly deep in the Pacific, Craig, Clarke and their colleagues had calculated that the amount of helium 3 required to maintain the anomaly worldwide against the loss of helium to the atmosphere as the oceans overturn is some 1,100 moles per year. If all the helium 3 entering the oceans enters by way of ridge-axis hot springs, and if the ratio of helium 3 to heat is everywhere the same as the one measured at the Galápagos ridge axis, the worldwide transport of heat at the ridge-axis hot springs is readily calculated. It is 5×10^{19} calories per year, which places it in the middle of the range the geophysicists had calculated.

One could feel justified, therefore, in supposing that anomalies in the concentration of other elements in our samples would serve (as proportions of the helium-3 anomaly) to calculate the total fluxes of these elements from submarine hot springs into the oceans. The results are impressive, particularly when they are compared with the fluxes from continental weathering. Thus the ridge axes consume most of the magnesium and most of the sulfate that rivers introduce into the sea. Conversely, they release between five and 10 times more lithium and rubidium, and between a third and a half as much potassium, calcium, barium and silica. Their release of manganese is sufficient to account for all the accumulation of that element in the metalliferous sediments and nodules carpeting the ocean. Finally, the ridge axes convert back into carbon dioxide most of the bicarbonate produced by continental weathering. The low-temperature oceanic reactions Sillén envisioned are therefore replaced by a high-temperature hydrothermal process.



OPHIOLITE, a segment of oceanic crust emplaced on top of a continent, includes layers of sediment identical with the ones that carpet the ocean floor. It also contains sulfide ore bodies that are taken to have resulted on the ocean floor from the hydrothermal activity of a multitude of black smokers. Often the ore bodies occupy depressions in the top of a layer of basalt.

Black Smokers

It appeared from the analyses that conceptions of the processes controlling seawater chemistry over geologic time would need to be substantially modified. It would be well, of course, to discover hot end-member solutions emerging from the sea floor undiluted. The pervasive entrainment of groundwater at the Galápagos ridge axis made that prospect seem most doubtful. It was with surprise and excitement, therefore, that we heard news from Jean Francheteau of the Institut de Physique du Globe. The French research submarine *Cyana* had found large sulfide deposits on the ridge axis at the crest of the East Pacific Rise just south of the mouth of the Gulf of California at 21 degrees north latitude. The site gave no sign of current hydrothermal activity. Nevertheless, the deposits could have formed only by the precipitation of sulfides from a high-temperature solution.

We ourselves returned to the Galápagos ridge axis in the spring of 1979. Many more vent fields were found, but the water temperature was no greater than 23 degrees C. *Alvin* then left to take part in the work near the Gulf of California. Angus, already there, had photographed fields of hot springs, and in the midst of them *Alvin* found what we had thought would never be found. Great jets of black water spewed from chimneys of sulfide minerals that had built themselves up to heights of several meters. The temperature of the vent water exceeded 300 degrees C. (At the pressure of sea level, water at that temperature would boil explosively, as it does in geysers such as Old Faithful in Yellowstone. At the pressure of these depths, however, it remains liquid.) A scramble ensued to design and build water samplers that would function at temperatures exceeding 300 degrees. The National Science Foundation allocated additional time for *Alvin* to dive to the "black smokers." John A. Archuleta of the Los Alamos National Laboratory lent us equipment.

We dived in November, 1979. Navigation transponders left at the site guided us to the vents. We approached a black smoker issuing from the sea floor amid large blocks of sulfide ore. Maneu-

vering in a vigorous bottom current, the pilot inserted a newly designed temperature probe into the throat of the smoker, only 15 centimeters in diameter. The reading stabilized within a few tenths of a degree of 350 degrees. The predictive power of chemistry was vindicated! All the vents we sampled differed from a temperature of 350 degrees by no more than a few degrees.

The 350-degree solutions leave the vents as clear, homogeneous fluids. They contain 100 parts per million of iron and a few parts per million of zinc, copper and nickel. That represents enrichment by factors as great as 10^8; the concentration of iron, zinc, copper and nickel in ordinary seawater is measured in parts per trillion. In addition the solutions contain 210 parts per million of hydrogen sulfide; ordinary ocean water contains none. The concentration of silica is 1,290 parts per million, exactly what one would predict after examining the Galápagos data. In other words, the solutions are saturated with quartz. Conversely, the solutions contain no magnesium and no sulfate. That too is exactly what one would predict.

At the mouth of a vent the solutions quickly mix with the cold ambient seawater. The result is the black "smoke." It is a suspension of fine particles of iron sulfide that precipitate out of the solutions. The chimneys appear to grow by the accretion of a leading edge of calcium sulfate. The calcium is contributed by the hydrothermal solutions and the sulfate by seawater. As the leading edge advances it becomes exposed to the undiluted vent waters. Hence it redissolves, to be replaced by precipitated sulfide minerals. Vent water commonly breaks through leaks in the chimney, promoting lateral growth. The result is a complex multimineral deposit.

Exploiting the Ores

The vent solutions themselves would not be worth the effort to mine them. The recovery of a ton of zinc, for example, would call for the processing of about six million tons of solution, the amount that flows out of a typical black smoker over a period of several months. A ton of zinc is worth about $900. Plainly it is better to let nature take its course and exploit the resulting ore bodies.

First consider the black smokers. The formation of a massive sulfide deposit such as the deposits in ophiolites, which contain millions of tons of ore, would seem to require a forest of smokers. Perhaps the smokers originally occupy a depression in the sea floor; particles precipitating from the smoke might then fill the depression. In any event the hydrothermal deposit accumulating around the smokers will be reworked by hot solutions rising through it in their ascent toward the ocean floor. The deposit will

Which of these languages would you like to speak?

Mark the one you want to speak in 2 or 3 months' time

Afrikaans	☐	French	☐	Japanese	☐
American English	☐	German	☐	Norwegian	☐
Arabic	☐	Greek (Modern)	☐	Polish	☐
Chinese	☐	Hebrew (Modern)	☐	Portuguese	☐
Czech	☐	Hindi	☐	Russian	☐
Danish	☐	Indonesian	☐	Spanish	☐
Dutch	☐	Irish	☐	Swedish	☐
Finnish	☐	Italian	☐	Welsh	☐

A Linguaphone Course makes you feel at home in almost any country. You'll never miss the meaning of conversations or be at a loss for words.

- It must work—over 4 million Linguaphone students in 88 countries speak a second language FLUENTLY.
- Proven learning success. You LISTEN to real conversations on cassettes...UNDERSTAND what you hear by following illustrated textbooks. HOLD CONVERSATIONS with the speakers.
- **You start speaking the very first lesson.** You learn at your own convenience.
- It's like having a private tutor.
- You get a complete, professional language program at little cost.
- You gain a good, working vocabulary.
- In just 2 to 3 months you can speak another language with complete confidence.
- You develop an authentic accent. Only native-born speakers are used.

Linguaphone The Language Masters

MONEY BACK GUARANTEE—28 DAY FREE TRIAL

World Language Courses, Inc. Dept. 261 313 Nolana Ave. McAllen, Texas 78501

FREE INFORMATION: Please mail me FREE information about learning the languages I have checked. FREE brochure.

Name (please print) _____

Street _____

City _____ **State** _____ **Zip** _____

ANTIQUITY

A Periodical Review of Archaeology
edited by Glyn Daniel

Founded in 1927 by O. G. S. Crawford, ANTIQUITY has appeared regularly ever since and won acclaim the world over as the most authoritative journal in its field. While written by specialists, the articles, notes and reviews are popular in character and indispensable to all interested in the development of man and his past.

Professor Glyn Daniel, Faculty of Archaeology and Anthropology in the University of Cambridge, and Fellow of St John's College, has been Editor since 1956.

The annual subscription, postage included, is $35. Subscription forms and bankers' orders are available on request from

ANTIQUITY PUBLICATIONS LIMITED
Heffers Printers Ltd, King's Hedges Road,
Cambridge, England CB4 2PQ

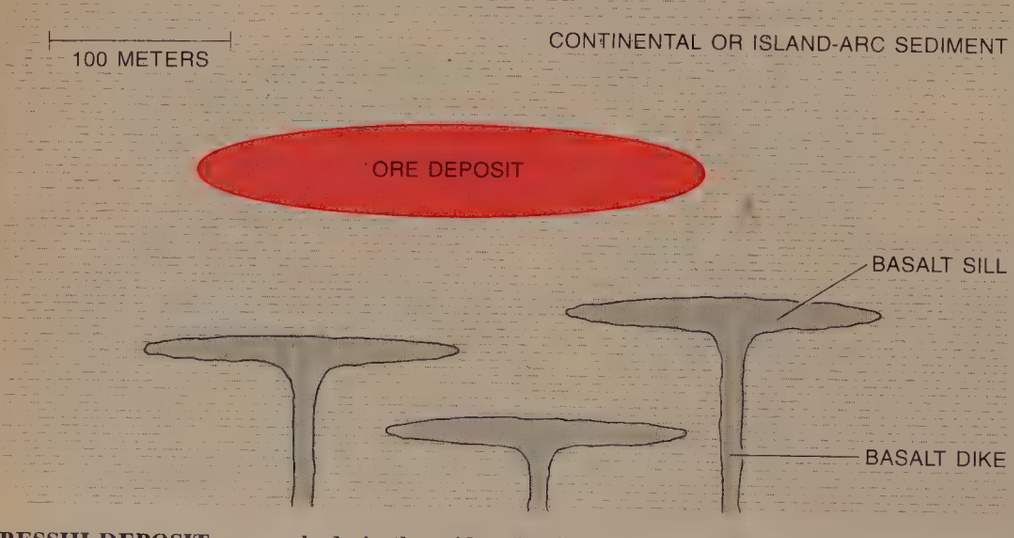
be stripped of its minor elements (copper, zinc, cobalt and nickel), leaving a matrix of large crystals of pure pyrite. The minor elements will be redeposited above the pyrite, in chimneys at the ocean floor. Meanwhile the leakage of cold seawater into the accumulating deposit will favor the in situ precipitation of fresh deposits from the ascending hydrothermal solutions.

This sequence accounts for the general features of the deposits that man has exploited in ophiolites. The main body of such a deposit is composed of pure, coarse-grained pyrite. The layer above the pyrite, which has oxidized to ocher, is enriched in the minor elements. Moreover, Alexander Malahoff of the National Oceanographic and Atmospheric Administration has recently discovered a forest of black smokers. About 30 kilometers from the Galápagos fields he observed from *Alvin* a profusion of extinct chimneys along a 600-meter length of the ridge axis. They were on top of a sulfide body several tens of meters high. Given the large amount of heat the chimneys must have released when they were active it is questionable whether they could have been approached at such a time by a research submarine. The Plexiglas portholes of *Alvin* soften at 86 degrees C.

It must be said that the deposition of ore at the site of black smokers themselves is extremely inefficient. An overwhelming fraction of the metals dissolved in the hydrothermal solutions become particles of the smoke that leaves the chimneys; this fraction is therefore dissipated by the currents flowing along the bottom of the oceans. Eventually the particles react with oxygen dissolved in seawater. The oxides of iron and manganese are insoluble in water; hence they form metalliferous sediments. The sediments include a certain amount of metallic copper, zinc, cobalt and nickel. Nevertheless, they are an ore too poor for commercial exploitation.

Precisely where is metal-rich sediment deposited? In the wake of Boström's discovery (with Peterson) that metalliferous sediments are ubiquitous on the ocean floor, he undertook to map the worldwide distribution of the sediments' metal content. On the floor of the South Pacific he found some peculiar patterns. At 15 degrees south latitude a long, narrow salient of sediment relatively rich in metal projected westward from the South Pacific ridge axis. At 30 degrees south latitude a broader salient projected eastward. Later John E. Lupton and Harmon Craig of the Scripps Institution mapped a spectacular plume of helium 3 extending westward from the ridge axis across much of the width of the Pacific at 15 degrees south latitude.

These various observations went without an explanation until a recent publication by Joseph L. Reid of Scripps. He had been surveying the density of the water from place to place in the oceans. From the densities he derives the shear forces between ocean layers. Then, having posited that the ocean has "surfaces" where water stands still (such surfaces are boundaries between stable currents flowing in different directions), he can calculate the magnitudes and directions of flow with respect to a given surface. Reid has developed maps of the motion of the water at a depth of 2,000 meters in the South Pacific with respect to a "level of no motion" that is taken to lie at 3,500 meters. Remarkably, the asymmetries found by Boström are exactly coincident with the flow lines inferred by Reid. The plume of helium 3 also stands explained. Helium is an unreactive chemical element; therefore it undergoes no change as it moves with ocean currents. In effect it labels the ocean water from which the hydrothermal effluent rains down on the ocean floor.



BESSHI DEPOSIT, an ore body in the midst of sedimentary rock, is thought to form under the ocean when a ridge axis close to a land mass is buried by erosional debris. Metal-rich sulfides aggregate hundreds of meters above basaltic dikes and sills: vertical and horizontal intrusions of magma that generate hydrothermal activity as they cool after entering the sediment.

Besshi Deposits

If the discoveries at the ridge axes are contributing greatly to the understanding of the processes that form ophiolite ore bodies, they are contributing with equal consequence to the understanding of a type of continental deposit of greater economic importance. It is the Besshi deposit, named after the one found in Japan. In a Besshi deposit a massive sulfide ore body lies in the midst of sediment, usually shale, a rock that originates as a fine-grained clay. The ore body does not lie on top of basalt. In fact, one must often drill in the shale to a depth of several hundred meters under the ore before any basalt is found, and even then the basalt takes the form of dikes and sills: vertical and horizontal intrusions of magma.

Evidently a Besshi deposit forms only when a ridge axis is close to a land mass that acts as a source of large amounts of erosional debris. One example is known in which the process is under way. At the Guaymas Basin in the middle of the Gulf of California the East Pacific Rise is penetrating into a continental plate. River sediment from Mexico has buried the axis in silt to a depth of several hundred meters. Under these circumstances the emplacement of newly minted oceanic crust takes an idiosyncratic course. The magma ascending at the site does not flow onto the ocean bottom. Instead it rises into the silt. The result is dikes and sills. The silt is quite permeable to water; hence hydrothermal activity is widespread. Indeed, the site was first identified (by Lawrence A. Lawver of Scripps) because of the great submarine flux of heat he had measured in the area.

Subsequently Peter F. Lonsdale of Scripps found sulfide mounds, each several tens of meters high and hundreds of meters long, at the center of the Guaymas Basin. The melting of the plastic liners of Lonsdale's core samples suggested a temperature of 100 degrees C. as little as 10 meters under the ocean floor. John Lupton of Scripps found high concentrations of helium 3 in the ocean above the site. Deep drilling penetrated hot basaltic sills.

In January of last year we took part in an expedition that Lonsdale led to the area. In seven dives *Alvin* sampled water venting from three of the largest mounds. The hottest waters had a temperature of 315 degrees C.; their very low concentration of magnesium and sulfate identified them as the end member of the system. In general, however, their composition made them dramatically different from the end members vented at ridge axes in the open ocean. Doubtless their ascent through the Guaymas silt is responsible for the dif-

ference. In particular the calcium carbonate shells of plankton buried in the silt are dissolved by the Guaymas end member, which as it leaves the basaltic reaction zone must be quite acidic (like the end members issuing from black smokers at 21 degrees north latitude). The addition of the carbonate, which makes the solution alkaline, induces the precipitation of sulfide minerals. By the time the Guaymas end member enters the ocean it is about four times more alkaline than seawater and is nearly depleted in iron and the other ore-forming elements.

The Smell of Diesel

At Guaymas, then, the crucial difference is really the richness of planktonic life in the overlying ocean. At Guaymas the end member encounters alkalinity under the ocean floor and so the metal content of the solution is quite efficiently trapped. On an open ridge axis, such as the one at 21 degrees north latitude, the end member encounters alkalinity only when it meets seawater. The result is black smoke and the dispersal of much of the end member's metal content.

At Guaymas the richness of plankton seems to have worked a curious modification of the vent-field ecosystem. When Lonsdale collected sediment cores in the Guaymas Basin, he noted that they smelled strongly like diesel fuel. When we returned last year, the water samplers on *Alvin* got clogged repeatedly with globules of wax. These findings are readily explained. The Guaymas silt is rich in planktonic carbon, which is "cracked" into hydrocarbons by the heat of the hydrothermal fields. The Guaymas fields are notable for great mats of bacteria; presumably the bacteria feed on the hydrocarbons. The bacteria appear to compete with clusters of tube-dwelling worms around the vents. Clams, mussels, fish and small crabs are missing. On the other hand, large horned crabs are common; they seem to prey on the worms. In all respects the Guaymas vent fields are different from the open ridge-axis fields as manifestations of hydrothermal activity. They hint at the variety to be expected in future discoveries.

The past two decades have seen an explosion in the earth and planetary sciences. On the earth the theory of plate tectonics has unified the hitherto fragmented and descriptive science of geology. Furthermore, the discovery of hydrothermal activity at the submarine ridge axes demonstrates that the new and the unique are not the exclusive province of spacecraft missions to other planets. It is said that the earth is the water planet. Now it can be said that the chemistry of the water and of ocean sediment bears the prominent imprint of volcanic processes.

Chivas Regal • 12 Years Old Worldwide • Blended Scotch Whisky • 86 Proof
• © 1982 General Wine & Spirits Co., N.Y.

A Window on the Sleeping Brain

REM (rapid eye movement) sleep, the phase of sleep when vivid dreams occur, is normally accompanied by paralysis. The paralysis can now be turned off in animals, making it possible to explore the REM phase

by Adrian R. Morrison

Sleep is ordinarily regarded as a condition of complete relaxation and inactivity. When the electrical activity of the brain in a sleeping human being or experimental animal is observed, however, it becomes evident that sleep is a complex and by no means inactive state. Indeed, the electrical activity of the brain in the phase of sleep when vivid dreams occur—REM (rapid eye movement) sleep—resembles that of wakefulness more closely than it does that of the other phases of sleep. In REM sleep transitory voltage changes indicate that the brain is in a state of arousal, in spite of the fact that sensory contact between the animal and its environment is much reduced. What prevents the neural arousal of the brain in REM sleep from being translated into vigorous physical movement is the absence of activity in the animal's muscles, which results in a paralysis that lasts until an episode of REM sleep ends.

What would happen if the paralysis that accompanies REM sleep were eliminated and the aroused brain were enabled to activate the muscles? Work in my laboratory at the University of Pennsylvania and in other laboratories is beginning to answer the question. The most primitive part of the brain is the brain stem, which lies between the spinal cord and the rest of the brain. In the cat small, precisely controlled lesions in the brain stem result in animals that show the electrical characteristics of REM sleep but move vigorously. The lesions are made in the region of the brain stem called the pons, and the exact position of the lesion within the pons strongly affects which muscles are freed from paralysis.

Therefore the lack of muscle tone in sleep appears to be directly attributable to the pons. Other experimental results suggest that neurons (nerve cells) in the pons also affect locomotion indirectly. In REM sleep one region of the pons appears to interfere with the action of a "locomotion center" that extends from the pons to a point farther down the brain stem. This interference prevents the locomotion center from activating the nerve networks in the spinal cord that are responsible for the reciprocal (alternating) motion of the limbs. Thus the activity of the pons in REM sleep leads to the inhibition of movement in two ways. My own work suggests that the link between heightened brain activity and muscular paralysis can also operate in wakefulness and may be the cause of certain sleep disorders.

The transition from alert wakefulness to sleep entails many changes in an animal's physiology and posture; the alterations collectively result in a profound change of state. Much of my work has been done with cats, and the changes described here apply specifically to the cat, but many of them are shared by other mammalian species. When a cat falls asleep, it assumes a familiar curled-up and relaxed pose. Its eyes close, and the nictitating membrane (the "third eyelid") covers part of the eye under the outer eyelids. There is a reduction in the number of muscle fibers that are contracting and therefore a reduction in the degree of tone in the muscles. In the initial phase of sleep the movements of the eye that can be detected in the waking animal disappear. There is a gradual lowering of brain temperature, generally by a fraction of a degree Celsius.

Underlying such changes are the changes in the neural activity of the brain that occur when sleep begins. The most general information about the electrical activity of the brain is provided by the electroencephalogram (EEG). The EEG is a record of continuous changes in electric potential measured between electrodes affixed to the skull. The rapid changes in potential represent the sum of changes in voltage in the outer membrane of large numbers of brain neurons.

The EEG in an awake and alert cat shows waves with a relatively low amplitude and a relatively high frequency. As the cat curls up and goes to sleep there is a notable change. The low-amplitude waves of wakefulness are replaced gradually by high-amplitude, low-frequency waves. Because of the presence of the low-frequency pattern in the EEG, the light initial phase of sleep is often referred to as slow-wave sleep. As the animal enters slow-wave sleep there are also changes in electrical activity in specific areas of the upper and lower brain. Such changes are recorded by means of electrodes inserted into the appropriate brain structure.

For many years the slow-wave pattern was thought to prevail throughout sleep. In 1953, however, Eugene Aserinsky and Nathaniel Kleitman of the University of Chicago noted in human beings intervals of sleep when the EEG record reverted to the high-frequency, low-amplitude pattern of wakefulness. After the EEG pattern had changed, the sleepers' eyes were periodically observed to move rapidly in various directions; it is for this reason that the intervals have been designated rapid-eye-movement sleep. Periods of REM sleep are now known to recur regularly during sleep, alternating with longer periods of slow-wave sleep. The timing and the duration of the REM periods vary with the species. REM sleep comes about once every 90 minutes in human beings and about once every 25 minutes in cats. Each episode lasts for several minutes.

By waking sleepers immediately after an episode of rapid eye movement had ended Aserinsky and Kleitman were able to show that REM sleep is associated with vivid and intense dreams. The identification of the time in sleep when the sleeper dreams was the source of much of the initial excitement over the discovery of REM sleep. Later work has shown, however, that the physiological aspects of REM sleep are equally intriguing.

In addition to the essentially identical EEG patterns in REM sleep and in the waking state the rates of activity of neurons in most of the subunits of the brain are quite similar in the two states. One of the most striking similarities is found in the hippocampus, which is phy-

logenetically the oldest part of the cerebral cortex. Throughout much of both REM sleep and wakefulness there is a regular pattern of waves in the hippocampus at a rate of about seven per second. This pattern, the "theta rhythm," is quite different from the activity of the hippocampus in slow-wave sleep, which is less regular and shows spike-like waves. An additional similarity is in brain temperature: after decreasing part of a degree in slow-wave sleep the temperature of the brain increases to roughly the waking level during an episode of REM sleep.

Hence many of the characteristics of REM sleep are parallel to those of wakefulness and are quite different from the characteristics of slow-wave sleep. Remarkably, all the characteristics mentioned above are signs of activation of the reticular formation, or reticular activating system, in the core of the brain stem. In 1949 Giuseppe Moruzzi and Horace W. Magoun, working together at Northwestern University, demonstrated that the reticular formation is responsible for arousal. Thus in REM sleep the brain is highly aroused. Because of the juxtaposition of rest and arousal, REM sleep is sometimes called paradoxical sleep.

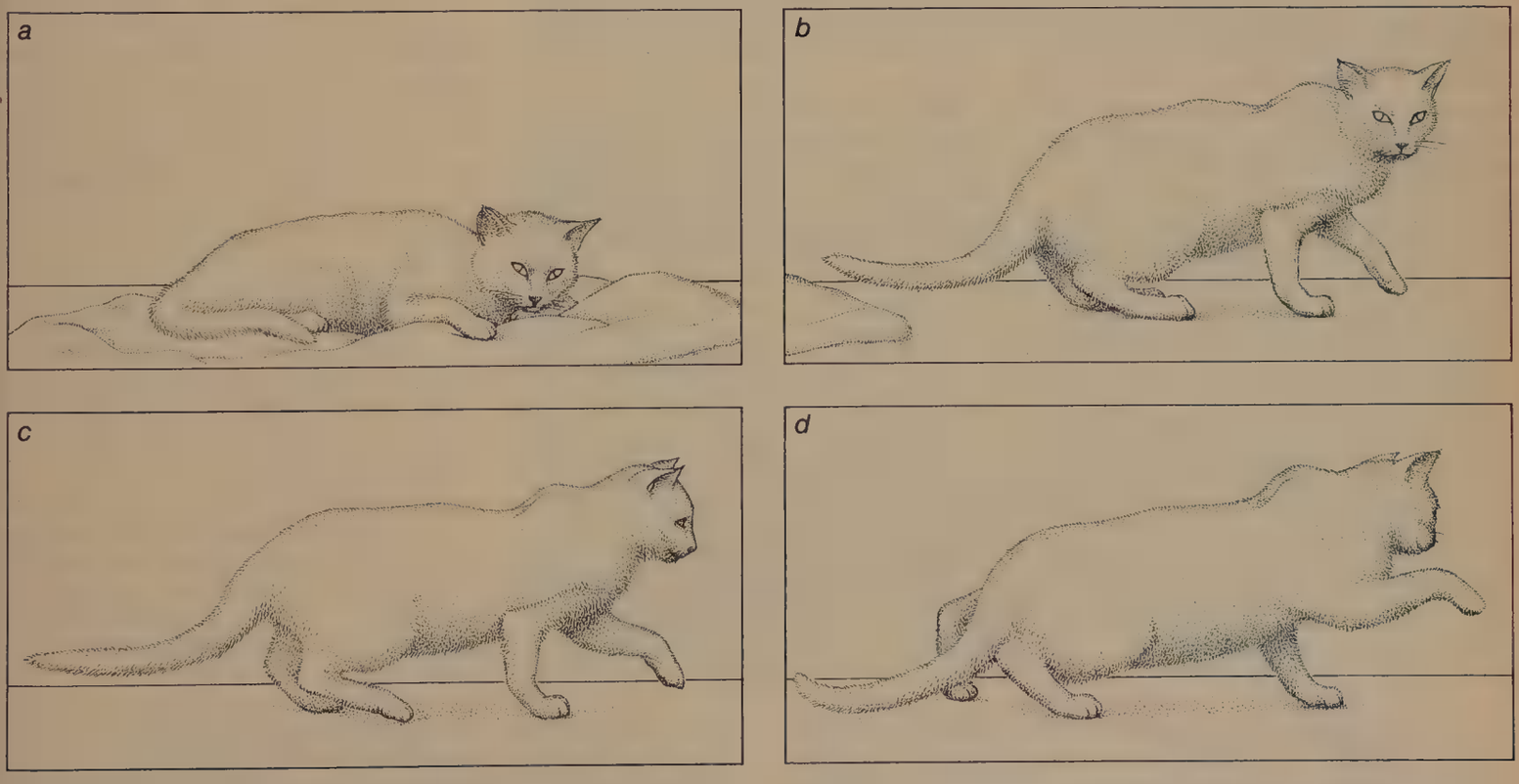
The close connection between REM sleep and alert wakefulness was emphasized recently by an unexpected finding in my laboratory. One of my students, Robert Bowker, had been working on electrical waves in REM sleep that appear "spontaneously" (meaning without apparent external stimulation). These brief, high-amplitude waves are called PGO spikes in reference to the brain structures in which they have been most studied: the pons (where they are thought to originate) and two parts of the visual system, the lateral geniculate body and the visual cortex (the occipital region of the cerebral cortex).

It had been thought that PGO spikes were limited almost exclusively to REM sleep and occurred only rarely in slow-wave sleep. One day in the laboratory, however, Bowker accidentally tapped on the recording cage while a record was being made from the brain of a cat in slow-wave sleep. Almost immediately a PGO spike appeared in the record. Further work showed that PGO spikes could be readily elicited in either REM sleep or slow-wave sleep by sounds and by tactile stimuli. Hence PGO spikes, which were thought to be spontaneous electrical events occurring only in REM sleep, were seen to be multisensory alerting responses that can be evoked in several brain states.

The conclusion that PGO spikes are alerting responses led Bowker to reexamine the waves called eye-movement potentials (EMP) that appear in the waking state. Such waves in the electrical record are identical in appearance with PGO spikes and are observed in the same brain structures. The presence of eye-movement potentials, however, was thought to depend on the level of illumination in the cat's environment. Such potentials had not been observed in awake cats in the dark. The dependence of eye-movement potentials on illumination was thought to constitute a significant difference from PGO spikes.

Reasoning that a cat in a dark recording chamber might be somewhat bored, Bowker aroused the animals' interest by sending the odor of tuna fish wafting through the cage. Soon the cat's record showed eye-movement potentials that were identical with PGO spikes. Short bursts of a loud tone yielded the same result. The result suggests that both PGO spikes and eye-movement potentials are particular forms of a general alerting response. The response can be elicited by stimulation in the waking state, in slow-wave sleep or in REM sleep. Activation of the same neurons can apparently also occur spontaneously in REM sleep. Thus in an episode of REM sleep the brain is aroused to the extent that it is in a state resembling alert wakefulness even though sensory contact with the environment is greatly reduced.

What prevents the animal in REM sleep, whose central nervous system is aroused and alert but for the most part cut off from sensory contact with the world, from injuring itself? The answer is that in each period of REM sleep the action of the motor neurons of the spinal cord that cause the skeletal (voluntary) muscles to contract is inhibited.



REM SLEEP WITHOUT PARALYSIS is shown in drawings based on motion pictures of a cat made in the author's laboratory at the University of Pennsylvania. The paralysis that normally accompanies REM sleep is eliminated by making a small lesion in the brain stem of the cat; the degree to which the paralysis is eliminated varies with the lesion. The cat in the filmed episode raised its head and righted itself during REM sleep (*a*). It managed to stand (*b*) and began to walk (*c*). It then began to paw at the air in a corner of the laboratory (*d*). In such episodes the cat shows all the signs of normal REM sleep, including the extension of the nictitating membrane (the "third eyelid") to cover much of the eye. That not all the muscle tone is released is suggested by the cat's slightly swayback appearance.

As a result the muscles are atonic, or without tone, and the animal is paralyzed. (It should be noted that atonia and paralysis are not identical. Paralysis can be the result of several conditions other than atonia. If the muscles are atonic, however, the animal will necessarily be paralyzed.)

The degree of activity in the skeletal muscles is often recorded by means of an electrode inserted into the neck. The neck muscles are among the "antigravity" muscles that are responsible for the maintenance of upright posture. As we have seen, when the cat enters light sleep, the tone in its muscles decreases because there is a decrease in the number of active fibers in each muscle. When the animal enters a period of REM sleep, the record of muscle tone goes flat, indicating a complete lack of tone. There are occasional bursts of muscle activity, however, that result in twitches in various parts of the body. In my laboratory we have been working with cats in which the atonia of REM sleep has been reversed by means of lesions in the pons made by destroying a small volume of tissue with a heated wire. The wire is introduced into the brain at predetermined coordinates.

After such a lesion is made, remarkable activity is sometimes observed. The activity follows a period of slow-wave sleep, at the time when the cat would ordinarily enter REM sleep with atonia. The cats raise their head, right their body and exhibit alternating movements of their limbs. They attempt to stand, and some succeed; others are actually able to walk. Cats with certain lesions display other behavior usually seen in wakefulness. They make motions typical of orienting toward prey, searching for prey and attacking. These movements are rarely directed at anything in the environment.

As I have noted, the kind of muscular activity the cat displays in these extraordinary episodes depends on the size and position of the brain lesion. When sections from the brain of a cat that could support itself on forelimbs only were observed with the light microscope, small, symmetrically placed lesions were noted in the dorsal, or upper, part of the pons. A cat that could support itself on all four limbs and walk was observed to have larger lesions in a more ventral, or lower, position. A cat that showed aggressive behavior, striking repeatedly at the floor in front of it, had lesions that extended forward into the midbrain. The significance of these positions will become clearer when we examine the neural pathways implicated in the inhibition of movement during REM sleep.

Several items of information suggest that the unusual movements we observe do indeed constitute REM sleep without atonia. The phenomenon was initially recognized by Michel Jouvet and François Delorme at the University of Lyons. As in ordinary REM sleep, the EEG of the cats that are active in sleep resembles that of wakefulness. The nictitating membrane partially covers the eye and the pupils close to slits. Recordings made from the hippocampus show the theta rhythm. Moreover, my student Joan Hendricks confined the sleeping cats in a padded harness, and at times when the cats with lesions would have been engaging in complex movements if they had been free they showed the rapid movements of eyes, whiskers and digits normally seen in REM sleep. Hendricks found that the temperature of the brain rose during these episodes, as it does in ordinary REM sleep.

Hendricks also showed that one particularly intriguing quality of REM sleep is present in REM sleep without atonia. Pier Luigi Parmiggiani and his colleagues at the University of Bologna demonstrated that the capacity of the cat to regulate its body temperature is suspended in REM sleep. In REM sleep cats appear to be "cold-blooded," like fishes, amphibians and reptiles. Hendricks showed that in REM sleep without atonia cats do not shiver or fluff up their coat in response to cold, although the same animals do show such responses when they are awake. Thus the lesions in the pons do not restore the capacity to respond to cold.

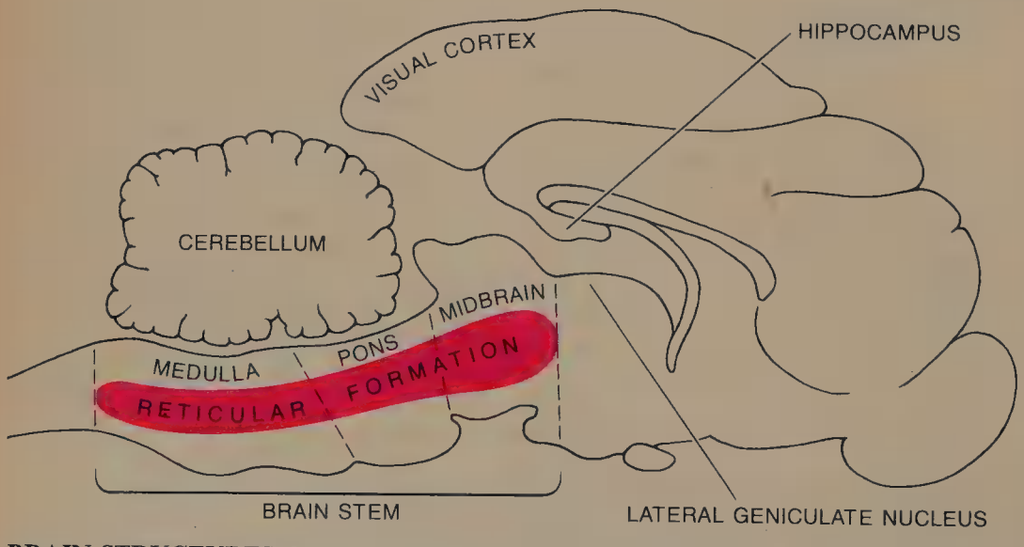
Much of my current work is aimed at understanding how the small lesions in the pons cause the atonia of REM sleep to be reversed and lead to elaborate movements that appear to be much like those of a waking animal. We have concentrated on the pons because in 1962, early in his work on sleep, Jouvet demonstrated that cats deprived of the entire brain in front of the pons still showed periods of atonia and rapid eye movements that were identical with those of REM sleep. In the course of my work I have concluded that there are at least two separate neural systems operating in REM sleep without atonia. One system is responsible for the release of muscle tone, the other for the liberation of "motor drive," a generalized impulse toward locomotion.

The first system appears to be the simpler of the two. It is known that an inhibitory center in the medulla is operative in REM sleep. (The action of neurons can be either excitatory or inhibitory. Inhibitory areas reduce the excitability of the target neurons.) In REM sleep the inhibitory center in the medulla interferes with the spinal neurons that activate the skeletal muscles. This mechanism is thought to be responsible for the lack of tone in the skeletal muscles.

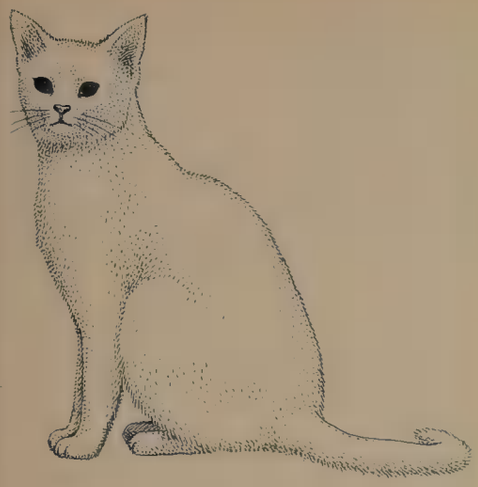
If the inhibitory center in the medulla were in turn under the excitatory control of a network of nerve cells in the pons, damage to the pons could break the excitatory connection. The action of the inhibitory center in the medulla would be interrupted and the skeletal muscles would retain their tone in REM sleep. The extent of the release of tone, and hence the muscles the cats could employ, would depend on the position of the lesions.

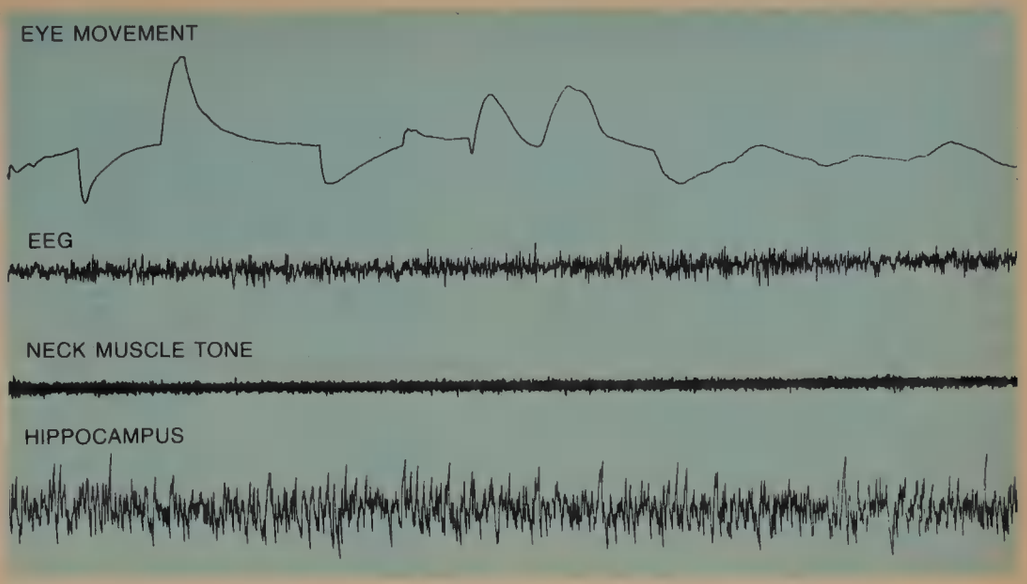
This hypothesis has received support from results obtained by Jouvet and his colleagues at Lyons. They have shown that a tract of nerve fibers originates in an area just dorsal to the area in the pons where we made lesions. The tract leads directly to the inhibitory area in the medulla. We have made lesions in the origin and in the course of the tract. Such damage yields cats that have some neck muscle tone in REM sleep but can lift their head only a little. This suggests that other neurons, probably in the reticular formation, must be damaged to release muscle tone completely.

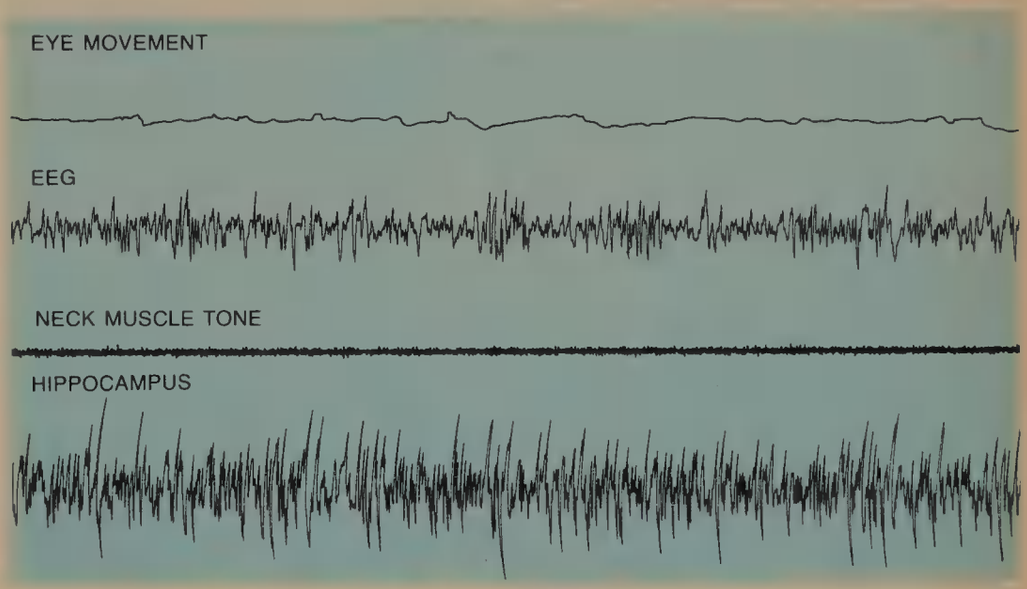
Therefore some experimental evi-

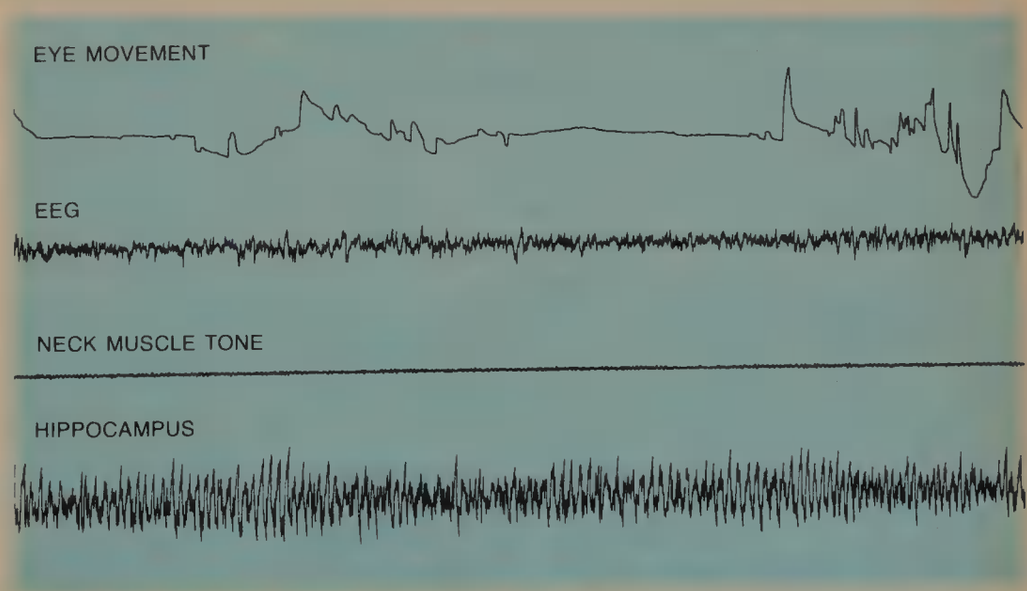


BRAIN STRUCTURES that play a role in sleep are indicated in a schematic section of the brain of a cat. The front of the brain is to the right. The lesions that eliminate paralysis in REM sleep are in the pons, a structure in the brain stem between the medulla and the midbrain.

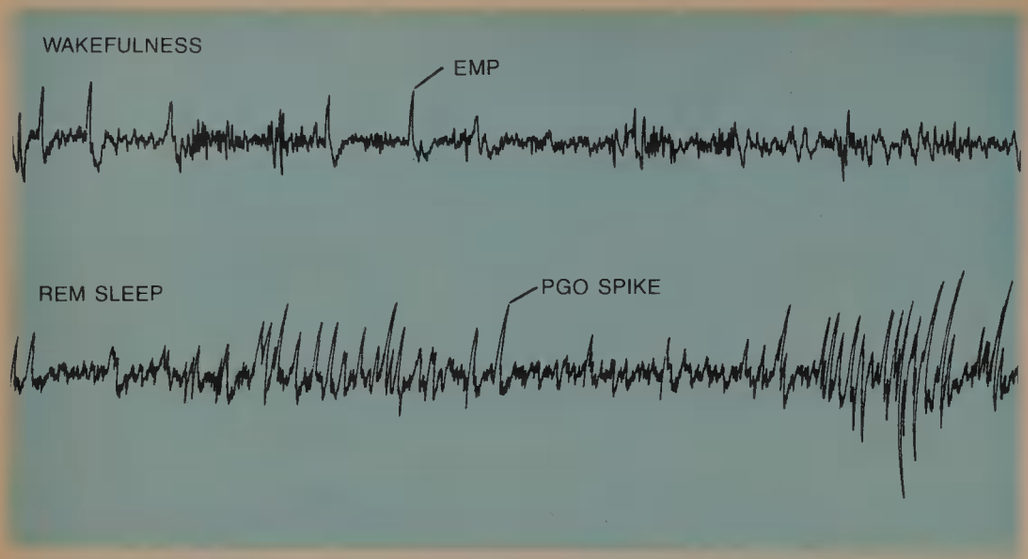




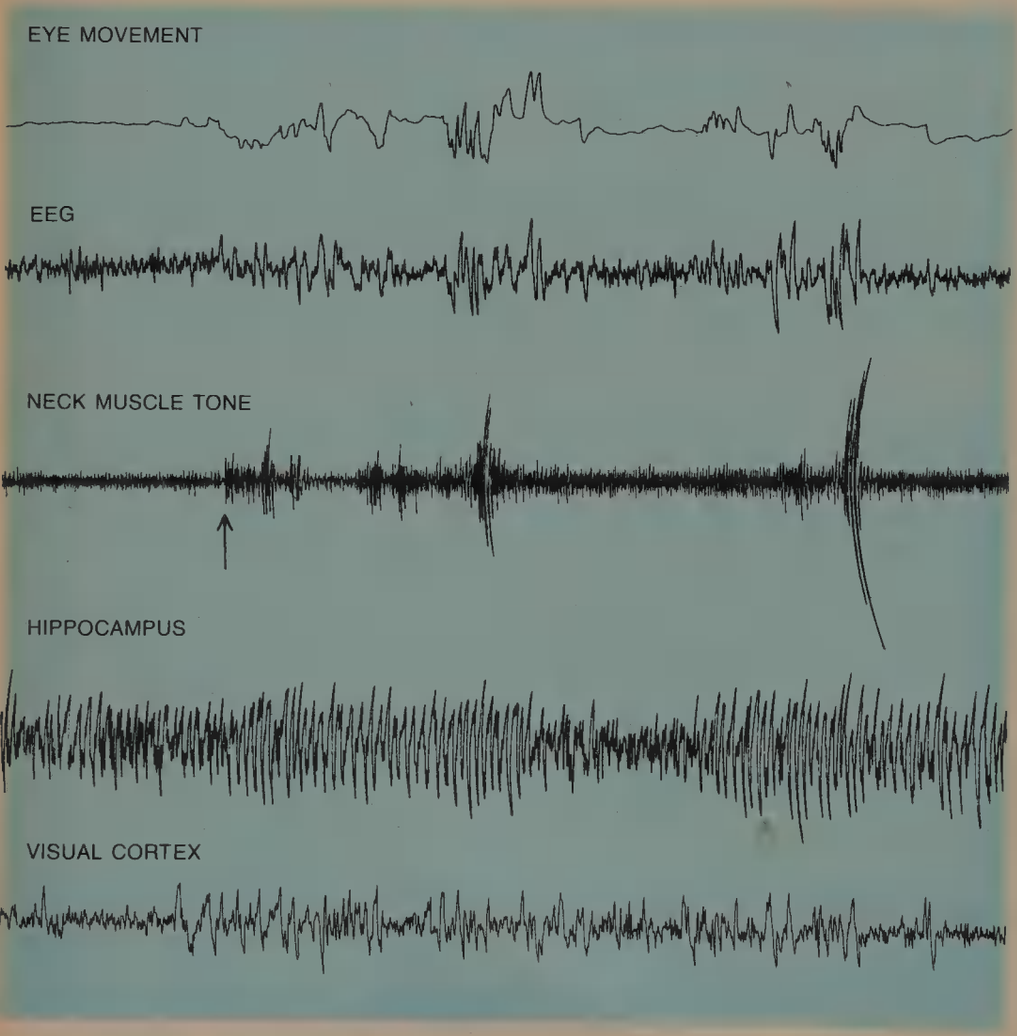




THREE STATES (wakefulness, slow-wave sleep and REM sleep) are detectable in the cat by changes in posture, muscle tone and brain activity. In wakefulness (*top*) the animal's eyes and head move as the animal responds to visual and auditory stimuli. The electroencephalogram (EEG), which records changes in voltage between electrodes put on the skull, shows a pattern of low-amplitude, high-frequency waves. There is substantial tone in the skeletal (voluntary) muscles, which is usually measured in the neck. The hippocampus, a primitive area of the cerebral cortex, shows a regular pattern called the theta rhythm when the cat is attentive to an object. As the animal enters slow-wave sleep (*middle*) it curls up. Eye movements cease. The EEG shows a low-frequency, high-amplitude pattern. Muscle tone decreases and the hippocampus shows an irregular rhythm. When the animal enters a period of REM sleep (*bottom*), which in the cat comes about once every 25 minutes during sleep, the curled-up posture relaxes slightly. Rapid eye movements, a high-frequency EEG pattern and the theta rhythm reappear. Skeletal muscle tone, however, disappears completely, a condition that is referred to as atonia.



BRAIN AROUSAL IN REM SLEEP was demonstrated recently in the author's laboratory. In the waking state waves called eye-movement potentials (EMP) appear in the visual area of the cerebral cortex (*top*). Such potentials had been thought to depend solely on the level of illumination. The author's student Robert Bowker showed that eye-movement potentials can be elicited by a variety of stimuli even in absolute darkness: they are multisensory alerting responses. Such waves are identical with the waves known as PGO spikes that appear spontaneously in the same areas of the brain in REM sleep (*bottom*). In REM sleep the brain is alert. The alertness, however, results from the brain's activity rather than from sensory information.



EPISODES OF MOVEMENT DURING SLEEP in the cats with lesions take place when all signs except muscle tone are similar to those of REM sleep. The similarities have led the author to conclude that the episodes of movement represent REM sleep without the usual atonia. Recordings from normal REM sleep include rapid eye movement, low-amplitude EEG pattern, theta rhythm and PGO spikes. All are present in recordings such as the one shown, made during the episodes with movement that follow damage to the pons. The notable difference is that in REM sleep without atonia the cat's skeletal muscles have regained much tone, enabling the animal to move. The arrow indicates the start of an episode of movement during REM sleep.

dence bears out the notion that a fairly direct connection between the pons, the inhibitory center in the medulla and the skeletal muscles underlies the atonia of REM sleep. Since REM sleep in human beings is known to be associated with intense dreams, it would be tempting to conclude that by breaking this connection we are enabled to witness the animal acting out its dreams. Apart from the difficulties inherent in attributing complex mental states to other species, however, there are good reasons to think this is not a full account of what happens in the episodes of REM sleep without atonia.

The functions of the brain are carried out by closely interconnected anatomical structures. It would be unreasonable to expect that the only result of damaging a central area of the brain stem would be to destroy structures inhibiting motor neurons. Indeed, we have evidence that systems other than the inhibitory one in the medulla are affected by the lesions in the pons.

The most important evidence concerns the overall level of locomotor activity when the cats with lesions are awake. The cats show no abnormal increase in muscle tone in slow-wave sleep or in wakefulness; the only effect on muscle tone is in REM sleep. Nevertheless, the cats appear to be more active in general than they were before the lesions were made. When they are loose in the laboratory, they make something of a nuisance of themselves, running here and there to investigate inconsequential things.

In order to confirm the impression of unusual locomotor activity we conducted open-field tests. Several cats were tested before and after the lesions were made. The cats were put in a room with a floor that had been marked off in squares. The number of squares each cat entered in a 30-minute period served as a measure of locomotor activity. All the cats that displayed REM sleep without atonia also demonstrated increased activity after the lesions. The increases ranged from 30 to 261 percent and were all significant in a statistical sense. This finding led me to conclude that the lesions in the pons affect a source of generalized motor drive. It is probably an anatomical system different from the one that affects muscle tone, since when the animals were awake, their muscle tone was not affected by the lesions.

Our work is beginning to yield a picture of how the second system of inhibition in sleep could work. My hypothesis of how the system operates borrows heavily from work done on the neural control of locomotion. It has been shown that the region of the brain stem we are interested in can regulate locomotion without any contribution from higher brain centers. Such regulation in-

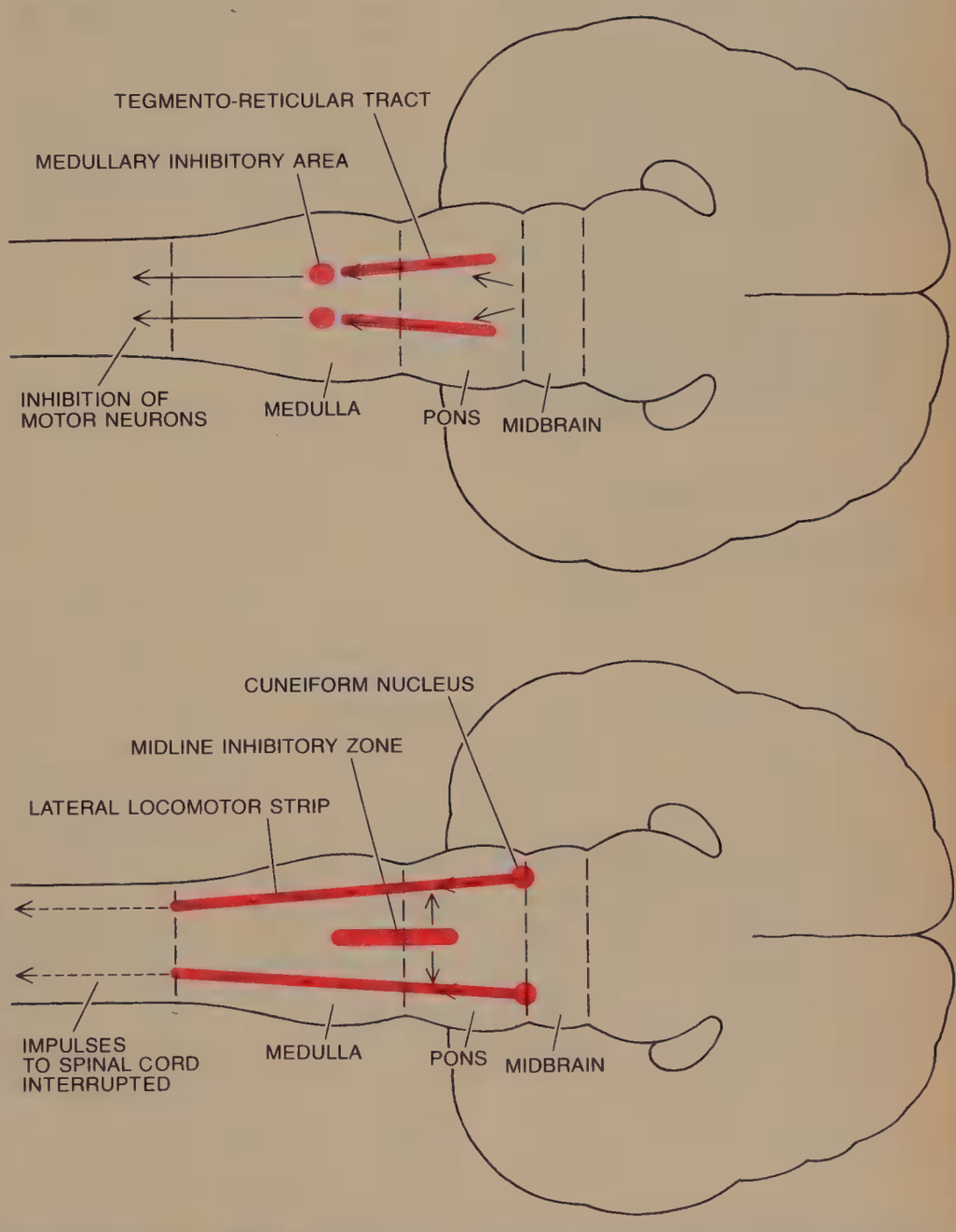
volves three structures. A "locomotion generator" in the spine includes neurons that control and coordinate the reciprocal motion of the limbs in walking and running. The generator is under the excitatory control of a second structure: a "locomotion center" in the brain stem. If the locomotion center is stimulated, it in turn stimulates the generator of reciprocal motion. The third structure is a control region in the pons, which is connected to the locomotion center by an inhibitory link. When the control region is activated, it suppresses the activity in the locomotion center. Thus in REM sleep the pons could indirectly suppress motor drive as well as muscle tone. Conversely, damage to the pons could release the muscles utilized in locomotion.

The existence of the brain-stem locomotion center and the spinal movement generator has been postulated by many workers, but their anatomical organization is not yet understood. The first evidence of the existence of the locomotion center was obtained by M. L. Shik, F. V. Severin and G. N. Orlovskii of the Moscow State University; their results have since been confirmed by other workers. Shik and his colleagues have shown that the application of an electric current to a neural system originating in the cuneiform nucleus, which is in the caudal region of the midbrain, can cause cats that have been deprived of the cerebrum to walk on a moving treadmill and even to trot or gallop. I suspect it is this center that the lesions in the pons release from inhibition in REM sleep.

Recent work by Shigemi Mori and his co-workers at the Ashikawa Medical College in Japan supports the existence of the locomotion center and suggests how the center may be connected to the pons. The Japanese workers have found that the tracts capable of inducing locomotion in cats deprived of the cerebrum extend from the midbrain down the outer part of the brain stem, in the form of two thin strips passing down the sides of the pons and the medulla. Moreover, they have shown that there are areas in the central part of the brain stem capable of facilitating or inhibiting the operation of the locomotion center.

Whether the locomotion center is excited or inhibited depends on the precise area that is stimulated. Electrical stimulation of the more ventral facilitatory zone is capable of inducing reciprocal movement of the animal's hind legs without any direct stimulation of the locomotion center itself. Dorsal stimulation reduces muscle tone and inhibits locomotion. The facilitatory and inhibitory zones discovered by Mori and his colleagues are near the midline of the brain stem and are close to the regions in the pons where our lesions are made. Thus the connection between the pons and the putative locomotion center is becoming clearer.

Such findings have implications for interpreting the effects of the lesions in the pons. For example, in the cat mentioned above that was able to stand and walk, the lesion was capable of cutting off the full effect of the midline inhibitory zone on the locomotion center. In the cat that could raise itself on its forelimbs only, the lesion was not capable of interrupting the full midline inhibitory effect. My associate Graziella Mann and I have recently found that additional damage to the inhibitory zone in a cat capable of raising itself on forelimbs only can enable the cat to walk on all fours in REM sleep without atonia. Nevertheless, the anatomical connections in the pons and between the locomotion center and the spinal generator remain speculative. Further work in this area is much needed. In the cats that show aggressive behavior there are additional complications. Such animals always have damage in the neural path-



SEPARATE NETWORKS of neurons (nerve cells) appear to be responsible for inhibiting muscle tone and locomotor drive in REM sleep. The action of neurons on their targets can be either excitatory or inhibitory. An area in the medulla that is known to inhibit motor neurons plays an important role in reducing muscle tone (*upper panel*). The tegmento-reticular tract connects the pons to the inhibitory center. In REM sleep the pons is activated, exciting the medullary center by this pathway and others. The medullary center inhibits the motor neurons and gives rise to atonia. A lateral locomotor strip plays an important part in the reduction of motor drive (*lower panel*). The strip runs down the outside of the brain stem. It is connected to structures in the spinal cord and can induce limb motion in cats from which the cerebrum in front of the midbrain has been removed. In the core of the brain stem are neurons that can excite or inhibit the locomotor strip. They are arranged in facilitatory and inhibitory zones, lying in the same vertical plane near the midline of the brain. In REM sleep the pons stimulates the inhibitory zone, turning off the locomotor strip and shutting down motor drive. Thus damage to the pons can eliminate inhibitory control in both systems, freeing muscle tone and motor drive.

ways descending from the cerebrum that are known to be involved in the control of aggression in wakefulness.

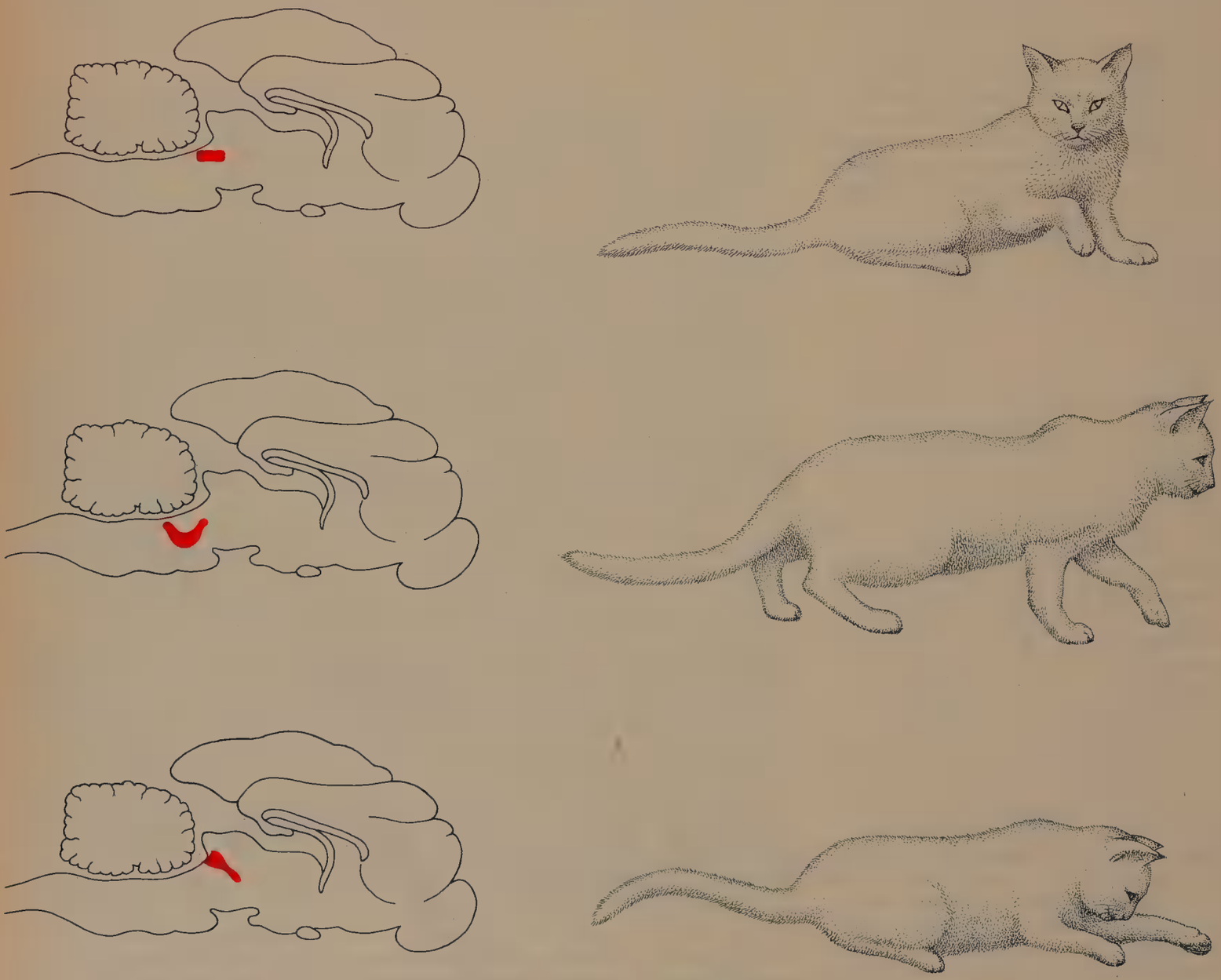
It should not be assumed, however, that even when the specific anatomical connections have been worked out, the control of the muscles in REM sleep will be completely understood. Even subtler mechanisms than the ones visualized here could be operating. For example, René Drucker-Colin of the University of Mexico and his students Gloria Arankowsky and Raul Aguilar and I have recently shown that the high-frequency neuron discharges occurring in many regions of the brain in REM sleep are necessary for the release of activity in REM sleep without atonia.

The administration of chloramphenicol, a common antibiotic, at a dose comparable to that employed in treating bacterial infection causes atonia to reappear in cats with lesions in the pons that would otherwise show episodes of movement in REM sleep. The cats appear to be normal when they are awake. Drucker-Colin had previously demonstrated that chloramphenicol reduces the rate of firing of the neurons in the reticular formation. The mechanism by which the neural activity is reduced is not known. Chloramphenicol, however, is known to be an inhibitor of protein synthesis, and this action could be responsible for the reduction in the firing rate of the neurons. Whatever the mechanism of the drug is, after it is administered the lesioned cats appear to have too little locomotor drive to generate the elaborate activity they would otherwise show in REM sleep.

The work that has been done so far on REM sleep without atonia indicates there is a link between the heightened arousal of REM sleep and the reduction of coordinated motor activity. Obviously an organism is well served by a mechanism that makes it impossible to move when the brain is very active but unresponsive to external stimulation. Teleological arguments need not be invoked, however, to explain the existence of a connection between brain arousal and motor inhibition. Such a connection could well exist even in the waking state.

Faced with a novel or unexpected stimulus, a person or an animal normally hesitates to some degree before acting. Most people have experienced a moment of hesitation or even a slack feeling in the knees when they have seen a fast-moving automobile bearing down on them. Such waking motor inhibition is generally short-lived. Nevertheless, it suggests that even in wakefulness there is a connection between momentarily heightened alertness and reduced motor activity.

The investigation of this link could have at least one important clinical consequence. The disorder called narcolep-



POSITION OF LESIONS in the brain of the cat strongly affects the kind of movement the animal shows in REM sleep without atonia. A cat with ■ small lesion high in the pons could support itself on its forelegs only (*top*). Such ■ lesion can interrupt only part of the effect of the midline inhibitory zone on the locomotor strip shown in the preceding illustration. A cat with a larger lesion deeper in the pons stood on all fours and walked (*middle*). Such ■ lesion can cut off the full effect of the inhibitory zone. A cat with a lesion extending into the midbrain displayed aggressive behavior (*bottom*). Such ■ lesion affects neural pathways from the cerebrum that regulate aggressive behavior.

sy is characterized by sudden and unpredictable lapses from wakefulness directly into REM sleep or from wakefulness into paralysis without loss of consciousness. Intriguingly, strong stimulation (such as that accompanying anger, laughter, surprise or sexual intercourse) are the commonest causes of narcoleptic seizure. It is possible that in those who suffer from narcolepsy there is an abnormally low threshold for the link between arousal and motor inhibition; common arousing stimuli could thus lead to atonia or to REM sleep in its entirety at inappropriate moments.

In addition to its possible clinical consequences the investigation of REM sleep without atonia will probably yield further information about the nature of REM sleep itself. It has already helped to clarify one central problem. As we have seen, in most regions of the brain the neurons follow the same pattern in the transition from being awake to being asleep. There is a decrease in the rate of activity from wakefulness to slow-wave sleep and then an increase to the rate of the waking state as an episode of REM sleep begins. Among the neurons that show a different pattern are those of the area of the pons known as the dorsal raphe nucleus. These cells utilize the substance serotonin as a transmitter to alter the activity of the neurons to which they are connected. The rate at which the cells fire decreases from about two impulses per second in wakefulness to almost zero in REM sleep after passing through an intermediate level in slow-wave sleep.

It had been suggested that the inactivity of the neurons in the dorsal raphe nucleus is a fundamental characteristic of REM sleep. Indeed, it had been hypothesized that the decrease in activity causes REM sleep to begin. To test these propositions I collaborated with Barry L. Jacobs and Michael Trulson of Princeton University. Jacobs and Trulson had already thoroughly studied the activity of the dorsal raphe neurons in the waking state and in normal sleep. When we recorded the activity of these cells in REM sleep without atonia, however, the results were surprising. In the cats with lesions in the pons the raphe neurons, instead of falling silent, increased their activity again after slow-wave sleep. The increase was to a rate of about one impulse per second and so did not bring the neurons to the rate in wakefulness. It did, however, yield a rate about six times that of normal REM sleep.

There are at least two plausible explanations for this unexpected finding. The first is that the unusual muscular activity in REM sleep without atonia somehow causes information to be relayed back to the pons to excite the neurons in the dorsal raphe nucleus. An alternative explanation is that the lesions in the pons affect a more central motor mechanism that normally inhibits the raphe neurons in REM sleep. The first hypothesis depends on the relaying of impulses from the periphery of the nervous system to its center whereas the second depends only on events in the brain.

To test the alternative explanations Jacobs and his students employed two drugs. The first is succinylcholine, a drug that is chemically closely related to curare. Succinylcholine acts at the junction between a motor neuron and the muscle cell that the neuron activates. By interrupting the connection the drug induces a temporary paralysis. When succinylcholine was injected into awake normal cats, the raphe neurons remained as active as they had been before the injection. Therefore the raphe neurons could be active in wakefulness with paralysis and in REM sleep without atonia and hence without paralysis. Since the raphe neurons were active in the awake, paralyzed animals, we concluded that the neurons of the raphe nucleus were not being stimulated in REM sleep without atonia by information relayed from the periphery as the result of vigorous muscular activity.

The second drug, carbachol, causes paralysis by directly affecting the mechanism described above that is responsible for atonia. The injection of carbachol silenced the raphe neurons almost completely. Rather than being a fundamental part of REM sleep or even an element in the mechanism that brings on an episode of REM sleep, the inactivity of the dorsal raphe neurons thus appears to be an ancillary phenomenon. It seems to be related to the central motor inhibition of REM sleep, which can be reversed by damage to the pons. Recently my student Peter Reiner has studied the other group of nerve cells that are turned off in REM sleep. These neurons lie near the site of our lesions and employ noradrenaline as a transmitter. Reiner's preliminary results suggest that the inactivity of these cells is also related to the motor inhibition of REM sleep.

By means of experimental techniques that enable us to separate the muscular paralysis associated with deep sleep from the activity that goes on in the brain we are learning much about what is fundamental to REM sleep and what is a side effect of REM-sleep episodes. The current hypotheses concerning the paralysis of REM sleep will undoubtedly require further work to be fully accepted, in particular the notion advanced above of separate pathways for the inhibition of muscle tone and of motor drive and the hypothesis that the second pathway is active in some way in wakefulness.

Whether or not these ideas prove to be sound, REM sleep without atonia

There's a lot worth saving in this country.

Today more Americans who value the best of yesterday are working to extend the life of a special legacy.

Saving and using old buildings, warehouses, depots, ships, urban waterfront areas, and even neighborhoods makes good sense. Preservation saves valuable energy and materials. We can also appreciate the artistry of these quality structures.

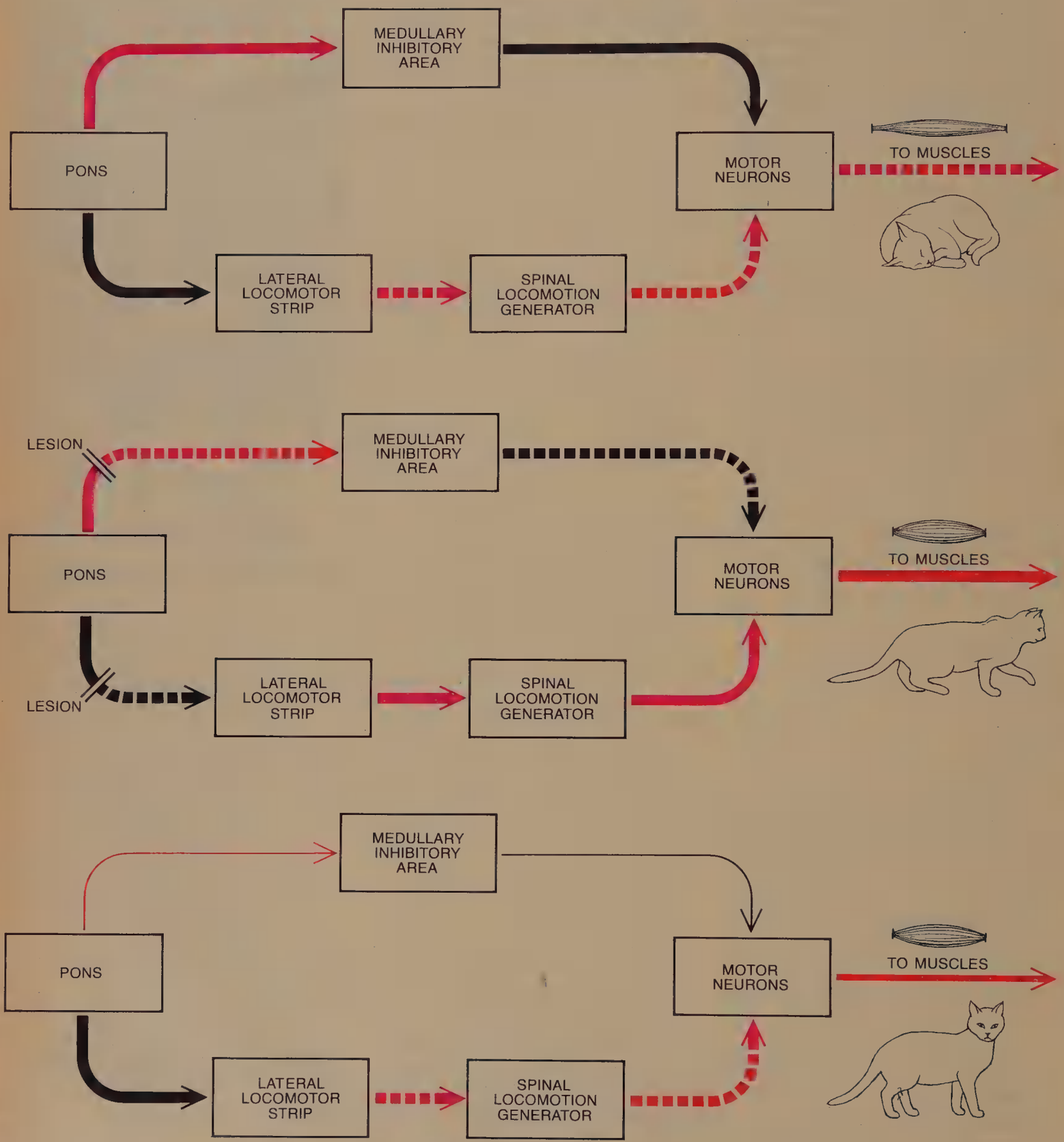
The National Trust for Historic Preservation is helping to keep our architectural heritage alive for us and for our children.

Help preserve what's worth saving in your community. Contact the National Trust, P.O. Box 2800, Washington, D.C. 20013.

National Trust for Historic Preservation

Preservation builds the nation

will continue to be a valuable experimental technique. In addition to making it possible to separate the essential from the ancillary in REM sleep, REM sleep without atonia could provide a means of studying complex behavior such as that of aggression, which is usually elicited in response to external stimuli but which could in REM sleep be generated in isolation from the animal's environment. Furthermore, because of the similarities between wakefulness and REM sleep it will undoubtedly continue to be interesting to compare the two states. In the movements of the waking animal we have a correlate of its inner states. Until now we have lacked such a correlate for REM sleep. In REM sleep without atonia we have acquired a correlate, which is in a sense a window on the sleeping brain.



SYSTEMS THAT INHIBIT MOVEMENT IN REM SLEEP may also operate in physical emergencies in wakefulness. The systems are shown in schematic form in the three panels of the illustration. Excitatory connections are shown in color and inhibitory connections in black. In normal REM sleep the pons strongly activates the inhibitory center in the medulla (*top*). The midline inhibitory zone in the pons inhibits the lateral locomotor strip. The result is complete paralysis. In REM sleep without paralysis the lesions break the connections from the pons to the locomotor strip and to the medullary center (*middle*). The cat can move, but the release of muscle tone is generally not complete. In wakefulness when an unexpected threatening stimulus is perceived, there is sometimes a transitory inhibition of movement (*bottom*). The temporary inhibition is probably the result of a reduction of motor drive, since there is little loss of muscle tone.

SCIENCE/SCOPE

Fusion energy machines that would turn sea water into electricity, though still 20 years away, are closer to fulfilling their promise of satisfying much of the world's energy needs. In plasma-heating experiments, Hughes Aircraft Company researchers have demonstrated the highest-performing gyrotron yet. It produced 285 kilowatts at 60 gigahertz at 45% efficiency under pulsed conditions. The short-range goal is to generate 200 KW at 60 GHz with long pulses in excess of 100 milliseconds. The long-range goal is 1 megawatt at 100 GHz. The Oak Ridge National Laboratory sponsors the program for the U.S. Department of Energy.

The electronic rocket engine is ready to be tested aboard a satellite to see how well it functions in the company of other space hardware. Hughes has delivered two engines, called mercury ion thrusters, for installation on a U.S. Air Force research satellite. The goal of the flight test is to qualify the system in space for performing such auxiliary propulsion functions as stationkeeping, attitude control, and orbit maneuvering of spacecraft. The system is designed to replace traditional chemical and gas propulsion systems, saving hundreds of pounds of weight. In operation, the thrusters are powered by the satellite's solar cells, which convert sunlight into electricity.

A new mobile radar automatically detects and tracks low-flying aircraft despite such severe clutter as a mountainous background. The Low Altitude Surveillance Radar (LASR) uses three-dimensional pulse doppler technology to pinpoint rapidly the location and altitude of ground-hugging aircraft. Current ground-based radars, with their broad-beam scanners, have trouble distinguishing a target from its background. LASR's pencil beams, which are much narrower, reduce clutter. Aircraft therefore can't hide in clutter for a surprise attack against front-line troops and armor. A prototype Hughes LASR tracked hovering helicopters and subsonic aircraft at altitudes from 10 to 6,000 feet.

Technologies of laser holography and diffraction optics have led to an experimental visor for protecting military pilots from potentially blinding laser beams. The visor reflects light at wavelengths used for lasers without significantly reducing visibility. It would replace devices employing dyes, which produce distracting discolorations, absorb light, and cut visibility. Designed by Hughes for the U.S. Navy, the visor could be adapted for ground troops.

A building-block family of electronic warfare equipment, which can be tailored for any class of Navy ship, automatically and instantly reacts to any threat of attack. The modular electronic warfare system (MEWS) offers electronic support measures (ESM), independent automated electronic countermeasures (ECM), or fully integrated ESM/ECM. MEWS tracks missiles and launching platforms, and provides jamming, high-powered deception, surveillance, and direction finding. Join our high-tech EW team. Send your resume to Hughes Ground Systems Group Employment, Dept. SE, P.O. Box 4275, Fullerton, CA 92634. Equal opportunity employer.

Creating a new world with electronics

HUGHES

HUGHES AIRCRAFT COMPANY

For more information please write:
P.O. Box 11803, Los Angeles, CA 90291

Early Farmers of the North European Plain

Excavations in Poland have revealed the remains of the people who farmed, herded and hunted in the area some 7,000 years ago. It was a time of transition, but in the end agriculture had come to stay

by Peter Bogucki and Ryszard Grygiel

One of the major events in human prehistory—the rise of agriculture and animal husbandry—took place at different times in different parts of the Old World. In southern Europe knowledge of this new way of life may have first diffused gradually from those parts of the Near East where the Neolithic Revolution had taken hold earlier. In Europe north of the Carpathians and the Alps, however, the Neolithic advance was a movement of people rather than ideas and one of the few such migrations clearly documented in the archaeological record.

Even so, after generations of field work not nearly enough is known about the people who left the valley of the middle Danube and carried the first cultivated plants, along with a number of domesticated animals, deep into the forested plain of northern Europe. The lack of data is due in part to the fact that the acid soils of the north quickly destroyed both human and animal remains and in part to the fact that archaeologists have preferred to work on larger and later Neolithic sites rather than on smaller and earlier ones. Nevertheless, in the 1930's early Neolithic sites on the Polish plain some 150 kilometers northwest of Warsaw had begun to yield valuable information until World War II halted the work. The same sites are continuing to yield information today. Before we describe them and their contents we should acquaint the reader with the physical nature of this broad transmontane belt and with the opportunities it offered to the first migrant farmers from the south.

Broadly considered, central Europe can be divided into three major ecological zones. The southernmost of them, the mountain areas of the Carpathians and the Alps, were generally not settled by farming communities until well into Neolithic times. Therefore that zone will not be discussed here. To the north beyond the mountains is a zone of rolling uplands extending from Slovakia and southern Poland through central Germany and into France. The region is largely covered with loess, the fertile but dry blanket of soil deposited by the winds that blew across Europe late in the Ice Age. The loess fills the river basins of the region and blankets the hills between them, reaching a depth of 100 meters in some places.

The third and northernmost zone consists of the flat lowlands of the north European plain, extending from the western U.S.S.R. across Poland, northern Germany and Denmark on into the Netherlands. The ice sheets of the last ice age covered this zone until some 15,000 years ago; their advance scraped out deep finger lakes and their retreat left a terrain of marshes, lake-filled hollows and a soil cover of sands, clays and gravels. Some of these new soils came to be very fertile; others remained marginal in fertility. All of them, however, supported a diverse and productive forest cover, and the lakes and streams were rich in aquatic life.

The farmers who entered the uplands from the south at first moved northwest along the major rivers of central Europe: the Danube, the Elbe, the Main and the Rhine. They settled in the richest upland ecological regions, the loess-filled river basins, where wild plants and animals were nonetheless relatively scarce. That left the newly established Neolithic communities largely dependent for their livelihood on their own imported animals (cattle, sheep, goats and pigs) and plants (barley and wheat). In time, however, some of the immigrants pushed farther north along the Vistula and Oder rivers and out onto the north European plain.

There they encountered terrain and soils considerably different from those of the uplands. The new ecosystem held more than 450 species of edible plants, fishes, turtles and mussels in the lakes and streams and red deer (*Cervus elaphus*), roe deer (*Capreolus capreolus*), wild cattle, wild pigs and even wild horses in the forest. This zone of tremendous natural productivity had only one drawback. Its yield was not constant throughout the year and was much reduced in the winter months. How the early Neolithic settlers adjusted to this unevenly productive ecosystem is one major question for which the Polish sites are yielding an answer.

Of a total of more than 20 Neolithic sites now known in the vicinity of Brześć Kujawski, a town in the Włocławek district of Poland, two (designated Site No. 3 and Site No. 4) were discovered in 1933 when farmers digging gravel near the shoreline of a dry lake bed came on artifacts and human bones. Konrad Jażdżewski of the State Archaeological Museum in Warsaw learned of the discovery and spent seven seasons (1933–39) at the sites, excavating a total area of more than 10,000 square meters. His work exposed the outlines of many Neolithic houses, more than 50 human burials, some of them richly furnished with copper, bone and bead ornaments, and antler pickaxes. Both human and animal bones were excellently preserved in the nonacidic soil. The copper artifacts were some of the earliest to be found in that part of central Europe; the metal had probably come from mines south of the Carpathians. The war halted Jażdżewski's research, and the artifacts uncovered in the 1939 season were destroyed when the Warsaw rail depot was bombed, but his work had already established the importance of the sites to an understanding of early Neolithic life on the north European plain.

Both Jażdżewski's work, done in the years before carbon-14 dating was known, and later research in the area left a number of chronological and economic questions unanswered. As a result one of us (Grygiel), of the Łódź Museum of

OUTLINE OF LONGHOUSE WALLS, exposed by the removal of plow-zone topsoil, guided the excavators at one early Neolithic site at Brześć Kujawski as they sectioned the discolored clay soil where the timber uprights had stood. Similar subsoil stains indicated the presence of trash pits filled with discarded artifacts and animal bones. The irregular excavation in the foreground led to a bed of clay the Neolithic settlers used to "plaster" their house walls and make pottery. This is the later of the two main settlements, established by people of the Lengyel culture. Calibrated carbon-14 dates indicate they came here between 4500 and 4400 B.C. and abandoned the site in 3900 B.C.

THREE BURIALS were associated with the longhouse remains seen in the top photograph. Skeletons are those of a man, at the left, lying on his right side, and two women, lying on their left side. The orientation of the bodies at burial placed the heads in a south-southeasterly direction. More than 70 Lengyel burials have been excavated but no graves of the first inhabitants of sites No. 3 and No. 4, people of the Linear Pottery culture, have been found so far. Many Lengyel burials are accompanied by rich grave goods, including copper ornaments.

Archaeology and Ethnography, accompanied by the other (Bogucki), then a graduate student at Harvard University, reconnoitered sites No. 3 and No. 4 in the summer of 1976. We found that substantial areas of both sites were still unexcavated. Receiving financial grants for further work from the Włocławek district and the town of Brześć Kujawski, we have now finished seven seasons of excavation, mainly at Site No. 4.

As at most Neolithic sites in temperate Europe, the archaeological remains here lie at a relatively shallow depth below the surface. The topsoil is a rich medium-weight Kujavian "black earth" that rests on a clay and gravel substrate. The upper 30 centimeters are regularly disturbed by plowing each year, and so the method of excavation is to strip away the plow zone, thereby exposing any features dug into the sterile clay and gravel underneath. Unlike the layered accumulation of deposits typical of a deeply stratified site, what comes into view is a palimpsest of pits, wall trenches, postholes and graves, which frequently overlapped. Such overlaps make it possible to determine the relative ages of successive settlements. The absolute ages can be determined by carbon-14 measurements; a total of 15 have now been made by laboratories in Poland, the Netherlands and the U.S.

Two main periods of Neolithic occupation are evident at the sites. They are separated by an interval of two to three centuries when settlement was on a much reduced scale. The initial period began in about 5300 B.C. (the true calendar date according to calibrated carbon-14 chronology) and continued to about 4800 B.C. The occupants belonged to a group assigned by European archaeologists to the Linear Pottery culture because of the distinctive incised-line decorations of their wares.

DEER-ANTLER SPOON, unearthed from one of the many trash pits at Site No. 4, is the first such artifact to be found among central European Neolithic remains. The Lengyel people also made pickaxes out of the antlers of red deer they collected after the animals had shed them.

MUSSEL-SHELL BEADS, strung in strands, were among the grave goods found with this Lengyel woman's skeleton. One such burial included more than 5,000 one-centimeter shell beads.

The remains of the culture are found all across Europe from the Ukraine to France, both on the loess uplands and on parts of the north European plain. Many of the upland sites show traces of sturdy longhouses, sometimes only a single "farmstead" but more commonly several dwellings grouped into a village. No Linear Pottery longhouses have been found on the north European plain, however, and the two sites we excavated are no exception. Among the large but shallow trash pits that indicated the Linear Pottery presence were only a few scattered postholes, probably marking the location of lean-tos or similar temporary shelters. Two separate phases of Linear Pottery occupation could be perceived. The earlier was at the tip of a lakeshore peninsula at Site No. 3 and the later was about 200 meters away at Site No. 4. No Linear Pottery burials were found at either site.

Centuries later emigrants assigned by European archaeologists to the Lengyel culture arrived to occupy both sites. This early Neolithic manifestation, which is named after the original site in Hungary, represents an essential continuation of Linear Pottery traditions in east to central Europe but is characterized by new pottery forms and a decline in the amount of decoration on the pots. Many Lengyel sites also have longhouses that are trapezoidal in plan, being significantly narrower at one end than at the other. Longhouses of this distinctive form have been found both in the loess uplands and on the north European plain; they are the first permanent shelters to appear at the two Brześć Kujawski sites.

Four Lengyel phases can be identified at the sites. The earliest phase, evident mainly at Site No. 4, appeared between 4500 and 4400 B.C. and continued until about 4300 B.C. The three later phases, which extended from about 4300 until about 3900, are evident at both sites. In the 1930's Jażdżewski named these later local forms of the Lengyel culture the Brześć Kujawski Group, a name that is still in use.

The most striking of all the features dug into the subsoil of the two sites are the traces of the Lengyel longhouses. They appear as long, dark discolorations in the clay and are the remains of trenches where the upright posts of the longhouse walls were set in place. Between the start of work in the 1930's and the present more than 50 such "house plans" have been uncovered, some of them more than 30 meters in length. All are oriented with their long axis running from northwest (the narrow end)

to southeast (the wide end). On the average the Lengyel longhouses were some 20 meters long; the northwest end was about three meters wide and the southeast end about five meters.

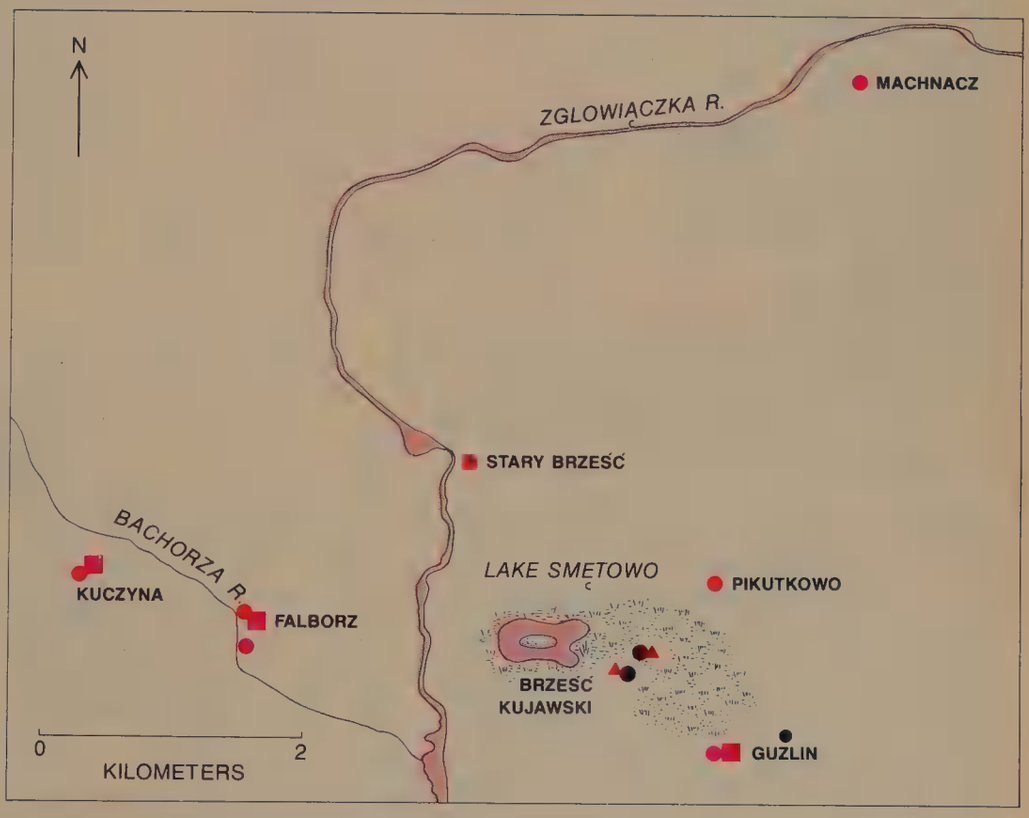
The reason for the Lengyel people's building trapezoidal dwellings remains uncertain, but a probable explanation is that the narrow end faced into the prevailing winds. In 1976 a replica of a symmetrical Neolithic longhouse was erected in France as a television prop. The end facing the prevailing winds has already suffered severely from weathering but the other three walls are still sound. Several instances of Lengyel house rebuilding are evident at the two sites; the fact that the same locations were often reused suggests the occupation was continuous through all four phases, probably with as many as 10 houses occupied at any one time.

The commonest and most informative subsurface features at the two sites are pits of various sizes and shapes. As we have noted, the pits associated with the Linear Pottery people's trash are generally shallow and large. At the time of occupation they may well have been nothing more than natural depressions in the ground, such as one left when a falling tree is uprooted. They contain dense accumulations of debris, consisting of animal bones and discarded stone, flint and pottery artifacts. In contrast, many of the Lengyel pits are deep, irregular holes dug in parts of the site where clay lies close to the surface. This suggests that the Lengyel people originally dug them to get clay for "plastering" their timber dwellings and for making pottery.

Other Lengyel pits are generally deep and circular and are found near the shore of the former lake. Silty deposits at the bottom of the pits suggest that at times they held standing water, probably for the maintenance of live mussels, turtles and fish. In any event most of the clay-yielding pits and the maintenance pits eventually served for trash disposal. Unlike the Linear Pottery pits, the Lengyel pits show considerable variation in their trash content. Animal bones and discarded artifacts are more numerous in the pits nearest the longhouses.

In some areas where the subsoil was disturbed in Neolithic times the discoloration is less than it is in the pits filled with trash; apparently the area that had been dug up had been refilled without any admixture of organic debris. On investigation such areas usually proved to be Lengyel graves. More than 70 Lengyel burials have now been excavated, the majority of them by Jażdżewski in the 1930's. The bodies were usually found in this kind of individual grave pit. Generally the skeletons are in a contracted position with the knees drawn up; males were interred lying on their

BRZEŚĆ KUJAWSKI SITES, indicated by colored triangle, are 150 kilometers west-northwest of Warsaw and about 200 kilometers northwest of the Holy Cross Mountains, an area where chocolate flint, named for its distinctive color, is found. This unique flint was prized by the Linear Pottery people because of its excellent flaking properties. Their successors at Brześć Kujawski were largely content with glacially deposited local flint of poorer quality.



SUCCESSIVE OCCUPATIONS at Brześć Kujawski stood on a peninsula jutting out into what was then a much larger Lake Smetowo. The Linear Pottery occupations at sites No. 3 and No. 4 were near the tip of the peninsula and near its base (*black dots*); the Lengyel occupations, at the same two sites, appear as colored triangles. Other Linear Pottery sites (*colored dots*) have been found to the south, west and north. Three nonresidential Lengyel sites (*colored squares*) have also been found. One of them, judging by the many axes found there, was a lumber camp.

right side and females lying on their left. With both sexes the horizontal orientation is the same: the skull rests near the south to the southeast edge of the pit. The females were often adorned with beads and other ornaments of shell, bone and copper; antler pickaxes are frequently found with the males. So far no Linear Pottery burials are known.

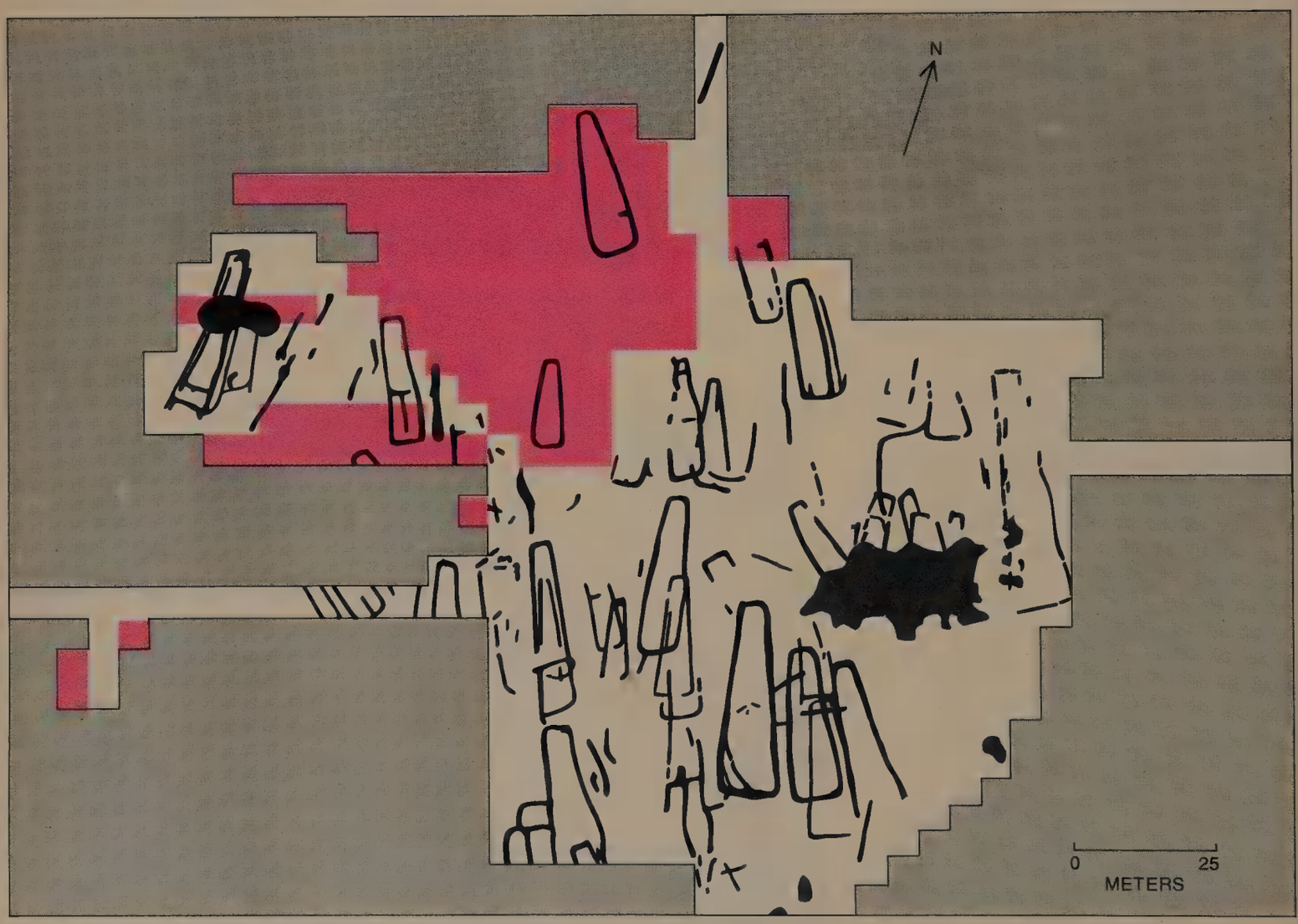
The assemblages of artifacts left by both the Lengyel and the Linear Pottery peoples include characteristically Neolithic axes and adzes made out of "pecked" and polished stone. Indeed, an ax of this type (known as the shoe-last celt) is characteristic of the Linear Pottery culture. At the same time these early farmers still made various small tools by flaking flint. The flintwork of the two cultures, however, was quite different. The Linear Pottery people favored a Polish specialty: chocolate flint, so called because of its brown color and found only in the Holy Cross Mountains some 200 kilometers to the southeast of the sites. Polish chocolate flint was valued by the makers of stone tools even in Paleolithic times. It has excellent flaking qualities, and the Linear Pottery people struck long, thin blades from chocolate-flint cores.

The Lengyel people showed no such strong preference. They made most of their flake tools from the locally plentiful flint "erratics" that the retreating ice sheet had left liberally scattered across the north European plain. Flint erratics, many of which have been subjected to pressure and other stresses in the course of being transported by the ice sheet, tend to have relatively poor flaking qualities, and the Lengyel flint flakes show it. Only when the Lengyel people needed a sharp precision tool did they take the trouble to procure and flake chocolate flint.

The Lengyel people went a step beyond the Linear Pottery people in making pecked and polished stone axes by drilling them so that a handle could be inserted; the stone plugs that were removed in the procedure are found in the trash pits. The Lengyel people also collected the shed antlers of red deer both to make pickaxes and to use as the raw material for other artifacts. Most notable among them is a handsome antler spoon, unique in central Europe, that we uncovered only last year.

Animal bones and teeth were also fashioned into beads and pendants. The ribs of large mammals (probably cattle) were bent into armlets, apparently while still "green," and then decoratively incised. Huge quantities of beads were made out of the pearly shell of the freshwater mussel (*Unio*). Strands wrapped around the waist of a female in one burial consisted of more than 5,000 *Unio* beads, each one about a centimeter in diameter. The graves of Lengyel females also contain copper beads and amulets; strips and bars of copper discarded in the trash pits indicate that these ornaments were made on the spot out of imported copper "blanks."

The pottery after which the Linear Pottery culture is named differs from the Lengyel pottery not only in decoration but also in form. At Brześć Kujawski the commonest shape is the same as the one most frequently encountered at other Linear Pottery sites in central Europe: a deep bowl forming three-quarters of a sphere. Another Linear Pottery shape that is common at the two Polish sites and elsewhere on the north European plain but relatively rare in the uplands is a sieve. We found no



MORE THAN 50 TRACES OF LONGHOUSE WALLS, many of them superposed, were excavated by Konrad Jażdżewski of the State Archaeological Museum in Warsaw at Site No. 4 between 1933 and 1939. The house outlines visible in the white area of this plan show the work of the 1930's. The dark gray areas indicate modern disturbance before Jażdżewski began work. Colored areas were excavated by the authors between 1976 and 1982. To avoid clutter burials and trash pits are omitted from the plan (*see illustration on opposite page*).

intact sieves, but the number of sieve fragments suggests that these artifacts served to separate curds from whey in making cheese. Robert Cowie of the University of Sheffield is currently conducting studies of Linear Pottery potsherds to see if their mineral composition offers clues to the sources of the clay used in their manufacture.

The Lengyel wares have shapes ranging from shallow bowls to deep vessels rather like Greek amphoras and includes such specialized items as pottery spoons. Their mineral content has not been subjected to detailed analysis, but the large amount of mica in the wares suggests they were made out of materials available locally. The same is suggested by the fact that clay was removed from pits at the site.

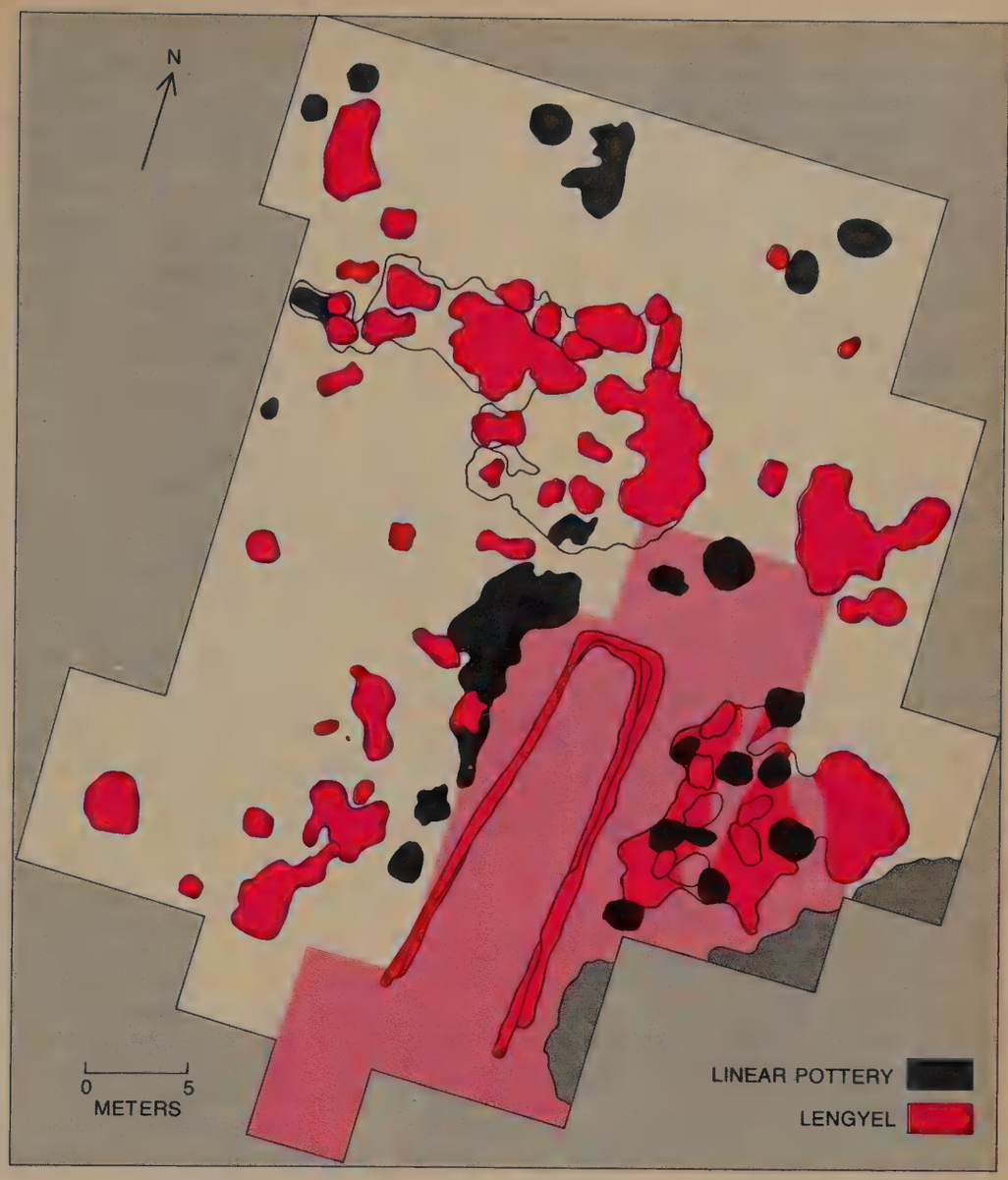
In our work since 1976 we have recovered a great many animal bones in a state of good to excellent preservation. For example, we have separated out more than 6,000 fishbones, mostly those of perch and bream. Here again there is a contrast between the animal remains characteristic of the earlier occupants of the site and those characteristic of the later occupants. As is the case with other sites of the culture on the north European plain, the primary domestic animals of the Linear Pottery culture were cattle. Since most of the cattle bones are at the small end of the overall size range, the herds appear to have consisted mainly of cows. Some sheep and goats were also tended by the Linear Pottery people, but whereas these farmers kept many pigs in their upland settlements, we have unearthed traces of only three pigs over the entire 500 years of the Linear Pottery occupation of Brześć Kujawski. Also notable for their small numbers are the remains of birds, fish, turtles and mussels and the bones of game animals: red deer, roe deer and wild pigs. Indeed, of the large number of fishbones found at the two sites, fewer than 100 come from the Linear Pottery occupation.

The animal bones associated with the Lengyel occupation tally differently. Although overall the remains of Lengyel cattle are somewhat more numerous than those of Linear Pottery cattle, these animals do not figure as prominently in the livestock totals. Sheep and goats are almost as numerous, and the pigs equal the cattle in numbers. Simultaneously there came an increase in the harvest of wild game: red deer and roe deer, waterfowl, turtles, fish and mussels. Not all the hunting was for meat. For example, beavers were evidently killed for their pelts, and so too, perhaps, were otters, weasels and hares.

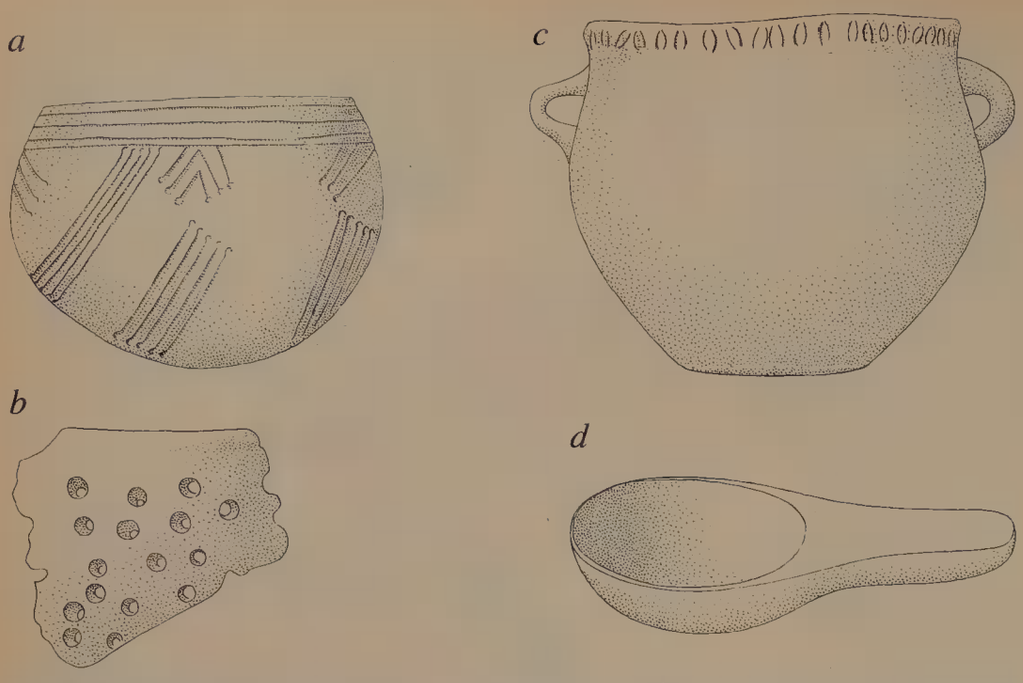
What about the tending of crops at these early Neolithic settlements? We have paid particular attention to the recovery of plant materials by means of wet-sieving and flotation techniques. We have succeeded in finding only three kernels of carbonized grain in the Linear Pottery trash pits, which suggests that these people engaged in very little grain cultivation. The Lengyel trash pits, in contrast, have yielded copious amounts of carbonized grain. One pit we excavated last year yielded more than 300 specimens. Preliminary analysis by Caroline Quillian Stubbs, who was then working at the University of Tübingen, indicates that most of the Lengyel grain is the early form of wheat known to botanists as emmer (*Triticum dicoccum*).

It is not clear whether the Lengyel people cleared forest to enlarge their farming activities. Such a conclusion would call for the analysis of pollen grains at various levels in the soil. No such pollen profiles are yet available from our two sites, although work to this end should soon be finished. One possible indication of forest-clearing activities, however, is the presence of the bones of eagles and marsh hawks among the animal remains. Both birds require open areas as part of their habitat.

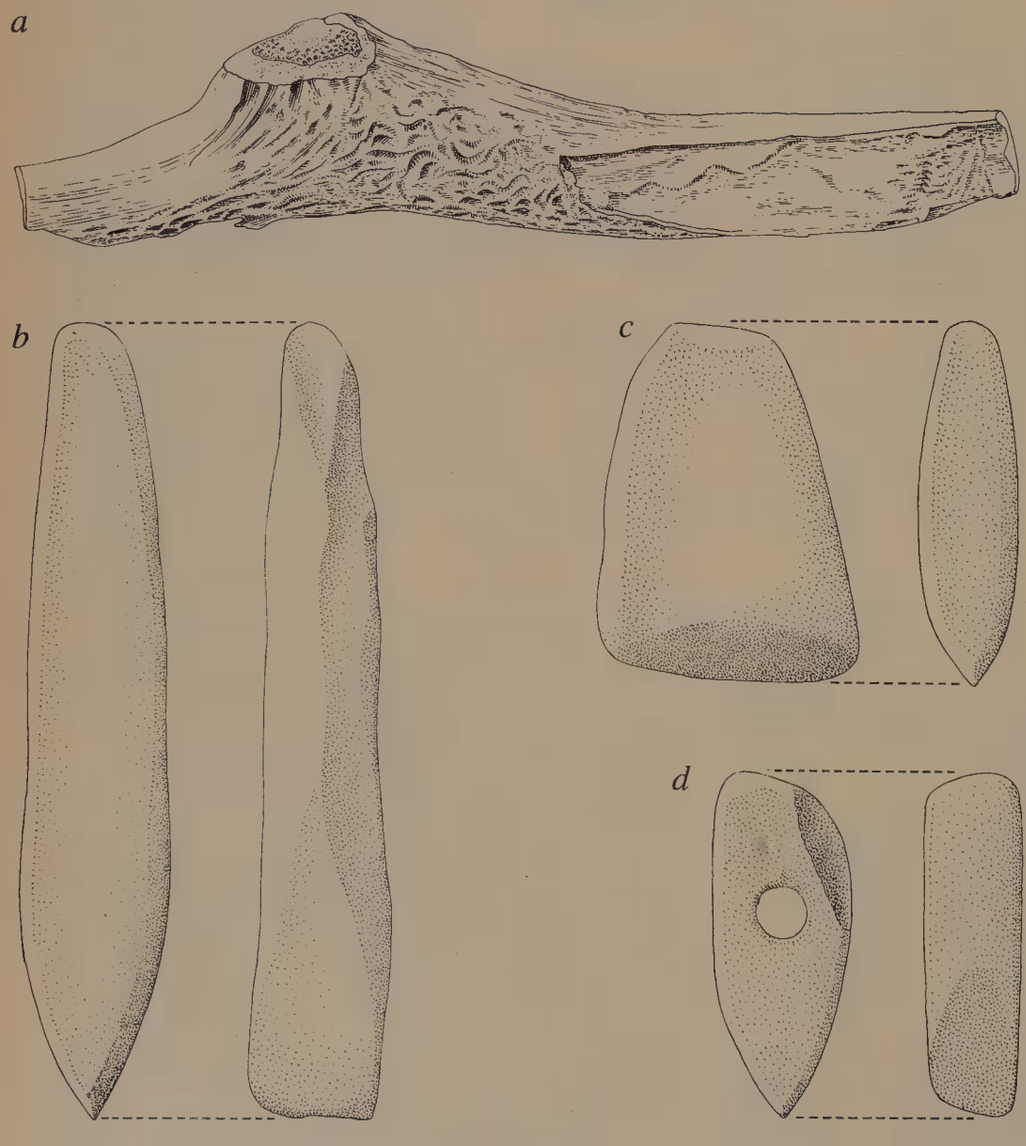
To determine what resources the early Neolithic populations exploited at various times of the year, we depended on a number of different clues, including the teeth of pigs and deer, and the degree to which the teeth still in the jawbones of the animals had erupted. We also noted other indications of seasonality such as the presence among the animal bones of the remains of certain waterfowl species (indicative of winter and spring hunting), of pond-turtle remains (the animals are inactive from late fall to early spring) and of shed deer antlers (they are cast annually in February and March). We could find no indication of winter activities in the Linear Pottery trash. At the same time the presence of turtle



WORK AT SITE NO. 3, nearer the lakeshore, done by Jażdżewski lies within the colored area of this plan. It included the exposure of two longhouse outlines almost perfectly superposed and a number of similarly overlapping trash pits that contained artifacts and animal remains of both the Linear Pottery and the Lengyel occupants of the site. The trapezoidal shape of the Lengyel longhouses is evident; the narrower end may have faced into the prevailing wind. Although the longhouses of the Linear Pottery people are found elsewhere in Europe, none are known on the north European plain. Postholes associated with Linear Pottery artifacts here suggest that these earliest Neolithic migrants constructed shelters of only a temporary kind.



DEEP POTTERY BOWL shaped like a three-quarter sphere (*a*) displays the linear style of decoration that gives the Linear Pottery culture its name. Below it (*b*) is part of ■ sieve, ■ common Linear Pottery artifact at the Polish sites, although intact sieves have not been found. The sieves probably served to separate curds from whey in making cheese. The simply decorated pot with ears (*c*) is ■ typical Lengyel form. Below it (*d*) is another common clay object of Lengyel manufacture: ■ pottery spoon. Neither the sieve sherd nor the spoon is shown to scale.



AXES were made out of pecked and polished stone or red-deer antlers (*a*). The stone ■■ at the left (*b*), known ■■ ■ shoe-last celt, is ■ characteristic tool of the Linear Pottery people. The shorter, broader ■■ (*c*) is ■ characteristic Lengyel form. The Lengyel people also drilled holes for handles in some of their axes (*d*). No intact drilled axes have been found at sites No. 3 and No. 4, but the stone plugs removed in making the holes have been found in trash pits.

bones in it indicates that the Linear Pottery group was active at the sites in the summer months. In contrast the contents of Lengyel trash deposits demonstrate that these later occupants of the sites lived there the year round.

The examination of "growth rings" in the teeth of mammals proved to be particularly informative. These layers of cementum are deposited as opaque bands in summer and as translucent bands in winter. By sectioning a tooth and determining which kind of layer was last deposited one can broadly determine when in the year the animal died. In this way we found that the Lengyel people usually slaughtered their pigs in midwinter and hunted red deer and roe deer in late winter and spring.

In addition to our work at sites No. 3 and No. 4 we reconnoitered the nearby countryside in search of other early Neolithic sites. We found evidence of six other Linear Pottery occupations, all showing the same characteristics as the main sites near Brześć Kujawski: they lacked evidence of permanent dwellings and their rubbish pits were shallow. Three additional Lengyel sites have also been discovered, all within an hour's walk of the two main sites. They also lack evidence of any permanent dwellings and the few pits they contain generally yielded arrays of artifacts more limited in variety than those at the main sites. They give the impression of being areas where some specialized activity took place, perhaps crop tending or herding. One of them, at Kuczyna, appears to have been a lumbering camp: its trash pits yielded an inordinate number of stone and antler axheads.

In all, both the nearness of the Lengyel camps to the main settlement area and their lack of permanent structures suggest that the Lengyel farmers occupied a well-defined "territory," whereas the more extended scatter of Linear Pottery camps gives the opposite impression. This, together with the functional differentiation of the Lengyel satellite camps, indicates that the Lengyel farmers exploited their surroundings in a more organized fashion than their Linear Pottery predecessors.

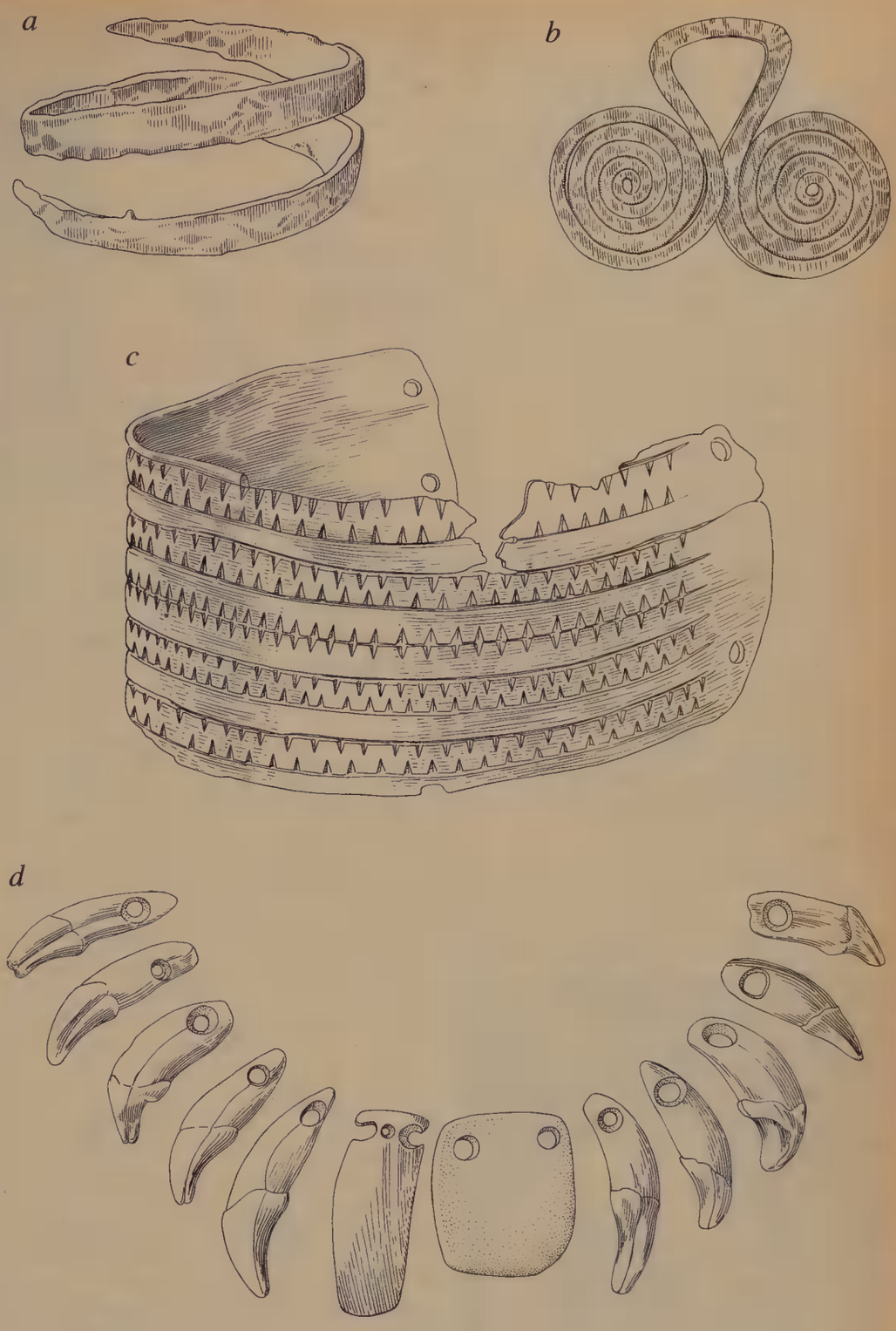
What are the general conclusions to be drawn from the work at Brześć Kujawski? In the past both the subsistence economy and the material culture of the first Neolithic communities in central Europe have often been implicitly characterized as "package deals." For example, wherever Linear Pottery is found it is presumed that an entire "package" of longhouses, pecked and polished "shoe last" axheads, grain, cattle, sheep, goats and pigs will also be present. At the two Polish sites, however, such is not the case, and neither is it at other Linear Pottery sites on the north European plain. The upland longhouses

are absent, the evidence of grain cultivation is sparse and pig bones are a rarity. Moreover, the Linear Pottery occupations of sites No. 3 and No. 4 show no indications of winter residence. The opposite is true of the Lengyel occupations of the same sites and also of Linear Pottery sites in the upland zone to the south.

We interpret this information as calling for a reconsideration of the way of life in the earliest food-producing communities of the north European plain. In view of the omissions from the "package" and the high proportion of cattle bones among the Linear Pottery animal remains we suggest that this first Neolithic excursion north of the uplands consisted not of permanent farming settlements but of seasonal cattle herding (and cheese making) in the lowland forests. The ecologically diverse vegetation of the lowlands would have provided rich summer forage for cattle in spite of the timber cover, probably far richer than the grazing available in the narrow valley bottoms of the uplands.

The conditions favorable for seasonal cattle herding may have been further enhanced in this forest area by an activity of the indigenous Mesolithic (pre-Neolithic) inhabitants of the area: the deliberate clearing of standing timber by fire. Evidence that such was a Mesolithic practice is provided by the abnormally high proportion of hazel pollen in the pollen profiles representing the period dated at about 5800 B.C., some 500 years before the first Linear Pottery people entered the lowlands. Hazel is a common second growth on cleared land.

Added evidence of the importance of dairy products to these earliest Neolithic immigrants is the presence of clay sieves both at Brześć Kujawski and at other Linear Pottery sites on the north European plain; these artifacts are comparatively rare at upland Linear Pottery sites. One may presume that one reason for it is that the upland settlers had not only grain but also numerous pigs, sheep and goats, whereas the lowland diet was limited to wild plant foods and dairy products. Another reason may be that the Linear Pottery people of Neolithic central Europe had not yet developed the tolerance to lactose (milk sugar) characteristic of modern European peoples; therefore they needed to process cows' milk into cheese (and perhaps yogurt) in order to consume dairy products without indigestion. Given the entire array of evidence, it seems probable that Linear Pottery herders came north with their cattle in the warm season but returned to permanent settlements in the uplands for the winter.



TYPICAL ORNAMENTS from Lengyel burials are a copper bracelet (*a*) and a copper "spectacle" pendant (*b*). Incised bone armlets (*c*) were made out of animal (probably cattle) ribs. The necklace (*d*) combines polished fossil-coral pendants with dogs' or wolves' teeth drilled for stringing. The copper artifacts, surprising in an early Neolithic context, were probably made on the spot out of copper carried here from mining areas south of the Carpathians.

The Lengyel farmers who succeeded the Linear Pottery people on the lowlands left behind evidence of the kind of diversification of resources that might be expected of a newly established year-round agricultural economy. Although they grew grain, it is doubtful that the emmer harvest was large enough to feed the community until the next harvest and at the same time provide seed for the spring planting. They must certainly have collected such wild produce as hazelnuts and goosefoot (*Chenopodium*). Goosefoot is a source of both greens and seeds that was widely exploited elsewhere in Neolithic Europe. It grows to a height of 1.5 meters in this part of Poland. The Lengyel people may also have supplemented their grain crop with such garden produce as peas and lentils; both are found at other Neolithic sites in central Europe, although no evidence of either has yet turned up at Brześć Kujawski.

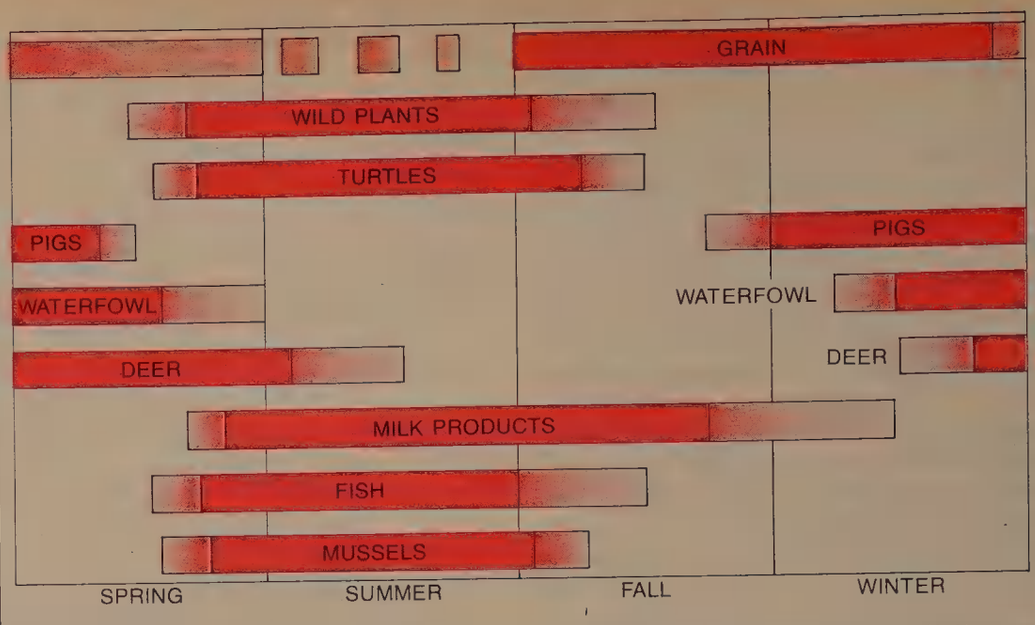
Of even greater importance in the Lengyel diet were the domestic animals the settlers herded and the wild game they hunted. As our tooth studies affirm, the Lengyel people slaughtered their pigs in winter after a summer of fattening, and they hunted red deer and roe deer for meat in late winter and spring. Turtles, fish and mussels would also have contributed to their diet in spring and summer, supplementing the wild

SCIENTIFIC AMERICAN

is now available to the blind and physically handicapped on cassette tapes.

All inquiries should be made directly to RECORDED PERIODICALS, Division of Volunteer Services for the Blind, 919 Walnut Street, 8th Floor, Philadelphia, PA 19107.

ONLY the blind or handicapped should apply for this service. There is a nominal charge.

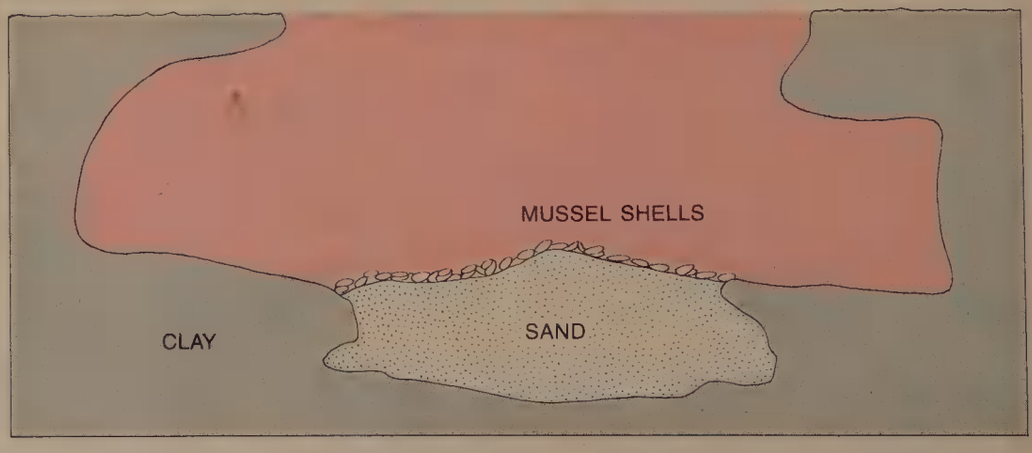


YEARLY FOOD CYCLE for the Lengyel people, beginning at the end of winter (*left*), saw ■ dependence on one domestic animal (the pig) and three wild animals (migratory waterfowl, red deer and roe deer) to bridge the gap until wild plant foods, freshwater mussels, turtles and fishes, together with milk from cattle, sheep and goats, became available in late spring and on into fall. With the wheat harvest in the fall the Lengyel people acquired a major resource that saw them well into the winter, when the summer-fattened pigs again became an important source of food, supplemented by waterfowl and deer and by cheese, ■ storable dairy product.

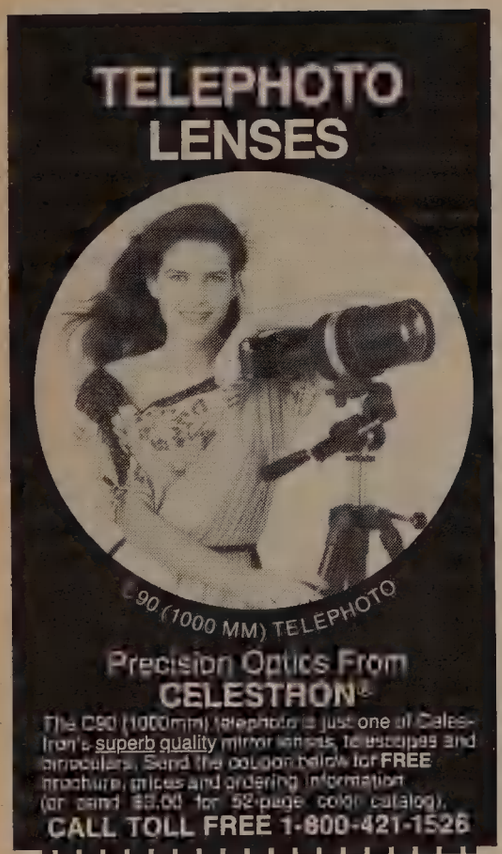
plant foods from the surrounding forest and the milk provided by cattle, sheep and goats. Such a combination of wild and domestic foodstuffs would have been particularly welcome in late summer, when the reaping and storing of the emmer crop would have called for a great deal of work in a short time.

The organizing and scheduling of this kind of subsistence system was evidently what made necessary a change in Lengyel times from the dispersed settlement pattern of the Linear Pottery people. We have mentioned the Lengyel satellite camps that contributed to the maintenance of the permanent residential area; there were probably also a number of other locales—hunting stations and herding outposts—that have left no trace in the archaeological record. The surrounding woods, fields and marshes must have been known intimately by the Lengyel farmers, every square meter contributing something to the well-being of the central settlement.

The long continuity of the Lengyel occupation testifies to the success with which these Neolithic farmers exploited the forest ecosystem around them. Any ecosystem, however, is resilient only up to a point. It seems likely that gradual environmental deterioration after hundreds of years of human occupation was what forced the abandonment of the Brześć Kujawski sites by about 3900 B.C. By then, however, food production as a way of life had become permanently established on the north European plain. The Neolithic Revolution was spreading along the Baltic coast and into neighboring northern Germany and Denmark in the hands of still later Neolithic peoples such as the people of the Funnel Beaker culture. The Linear Pottery and Lengyel pioneers were no more, but agriculture, with its innumerable consequences, had come to stay in temperate Europe.



PIT AT SITE NO. 3, close to the former lakeshore, is shown in simplified cross section. The presence of intrusive sand and ■ thin layer of mussel shells suggests that the pit was ■ "storage well," with standing lake water filling its bottom, used to keep mussels, turtles and fish alive.



CELESTRON INTERNATIONAL
P.O. Box 3578-SA, 2835 Columbia St.
Torrance, CA 90503

Name

Address

City State Zip

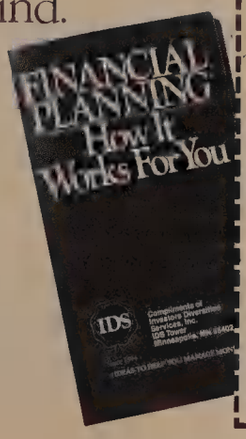
TAKE A LOAD OFF YOUR MIND WITH A FREE FINANCIAL PLANNING GUIDE FROM IDS.

Financial worries can spoil your appetite in a lot of ways. Interest rates fluctuate, tax laws change and inflation hasn't disappeared. Banks, brokers and insurance companies all seem to give conflicting advice. It's confusing. What you need is a carefully thought-out, completely objective plan that's flexible enough to work in the years ahead. And that's where IDS comes in. We're offering you a free guide to financial planning that explains what a financial plan is and how it can diminish your money worries.

An IDS representative can help you develop a personal financial plan to achieve your financial goals. You'll be talking to a career professional who understands the importance of financial planning. Someone who will regularly review your progress.

And because IDS offers over 50 different financial products and services, you'll get timely ideas to help you reach your goals.

Begin taking a load off your mind. Send for your free 24-page guide to personal financial planning. Or call toll-free 1-800-437-4332.



☐ **Yes, see that I receive my free 24-page guide to personal financial planning.**

Name_____

Address_____

City_____ State_____ Zip_____

Mail to: IDS, Dept. 583, IDS Tower, Minneapolis, MN 55402. Or call toll-free **1-800-IDS-IDEA,** that's **1-800-437-4332.**

60-52-376-191-001-C

IDS® **IDEAS TO HELP YOU MANAGE MONEY**

Chemical Signals of Social Amoebae

Two cellular slime molds can coexist in the same soil and yet maintain their identity. They do so by emitting and responding to different chemicals, both of which have now been identified

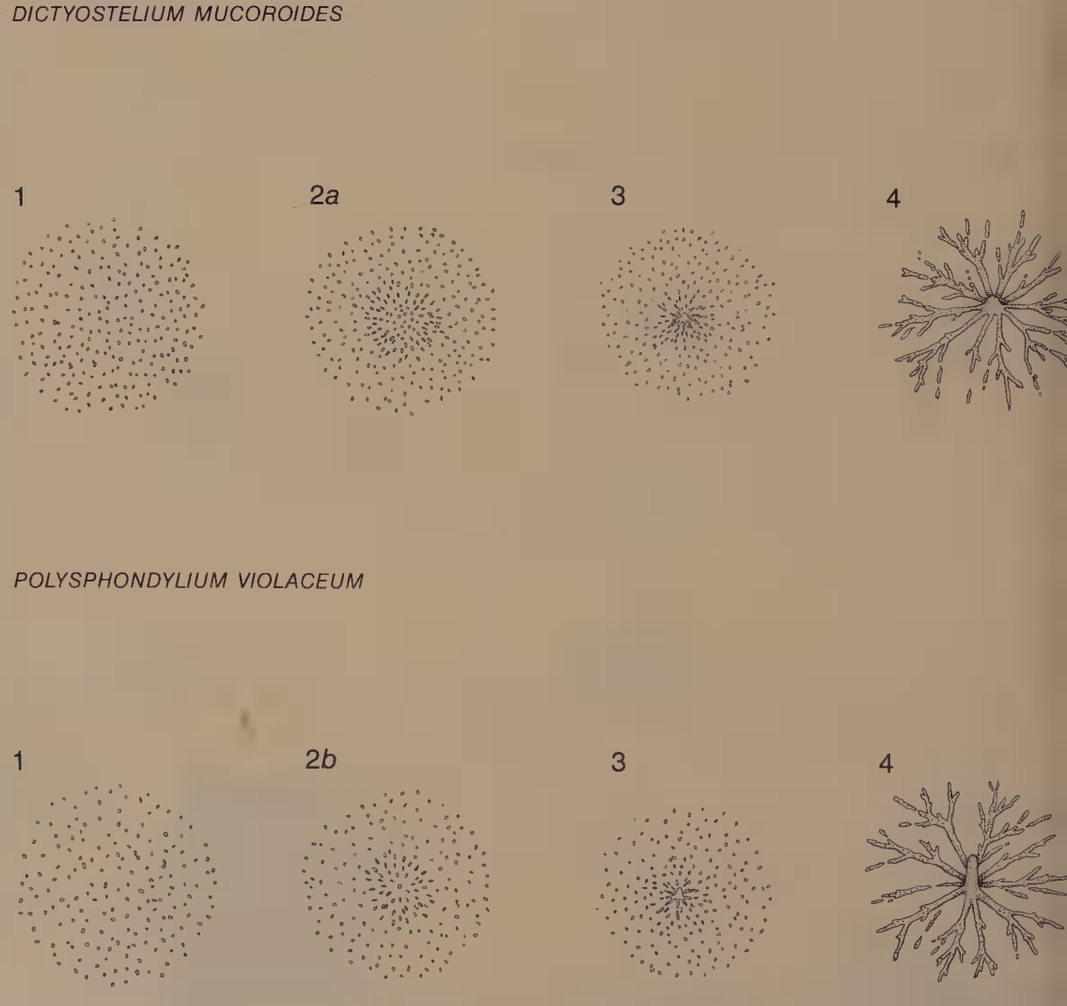
by John Tyler Bonner

In any bit of reasonably fertile soil two very small unicellular organisms are abundant: bacteria and amoebae. There may be millions of bacteria in a cubic centimeter of soil, and thousands of amoebae. The bacteria tend to be clumped near dead roots or dead insects, which supply them with nutriment, and the amoebae eat the bacteria. Each amoeba is about the size of a white blood cell, some 10 times longer than a bacterial cell. It crawls through the soil, engulfing bacteria it encounters, and every few hours it divides to form two daughter cells.

Prominent among the commonest kinds of soil amoebae are the social amoebae, otherwise known as the cellular slime molds. They feed as separate cells, like other amoebae, but once they have finished off all the bacteria in an area a trigger is set off within each cell that initiates their social phase. The cells stop their solitary wandering and begin to stream toward a central collecting point. Anywhere from a few hundred to 100,000 or so unicellular organisms come together to create a single, sluglike mass that has become a multicellular organism.

Then the cells begin to differentiate. A nipplelike tip appears at the front end of the mass, and inside the tip the beginning of a stalk takes shape. The stalk is a cellulose cylinder composed of swollen, rigid cells that ultimately die. The stalk builds upward as cells stream up the outside of it and pile onto the tip. When all movement has ceased, the mass of cells has become a fruiting body: a simple or branched stalk made up of dead cells encased in a cellulose sheath and capped by a little ball of spores. The spores are individually encapsulated amoebae, which become dispersed by water or by contact with a passing worm or insect. If the spore lands in a suitably warm and moist environment, it splits open, releasing a single amoeba that resumes the foraging phase of the slime-mold life cycle.

What I want to discuss in this article is a particular aspect of the social amoebae's behavior. Often there are two or more species of social amoebae in the same bit of soil. When they aggregate, they aggregate separately: the cells of different species become oriented toward separate collection centers. The streams of cells heading for one center pass through, and ignore, streams headed for another center, so that each species comes together at its own collection center and differentiates to form its characteristic fruiting body. How does it happen? Pieces of the answer have been emerging for some 40 years, and now it is finally possible to describe the segregation of the two species at the molecular level.



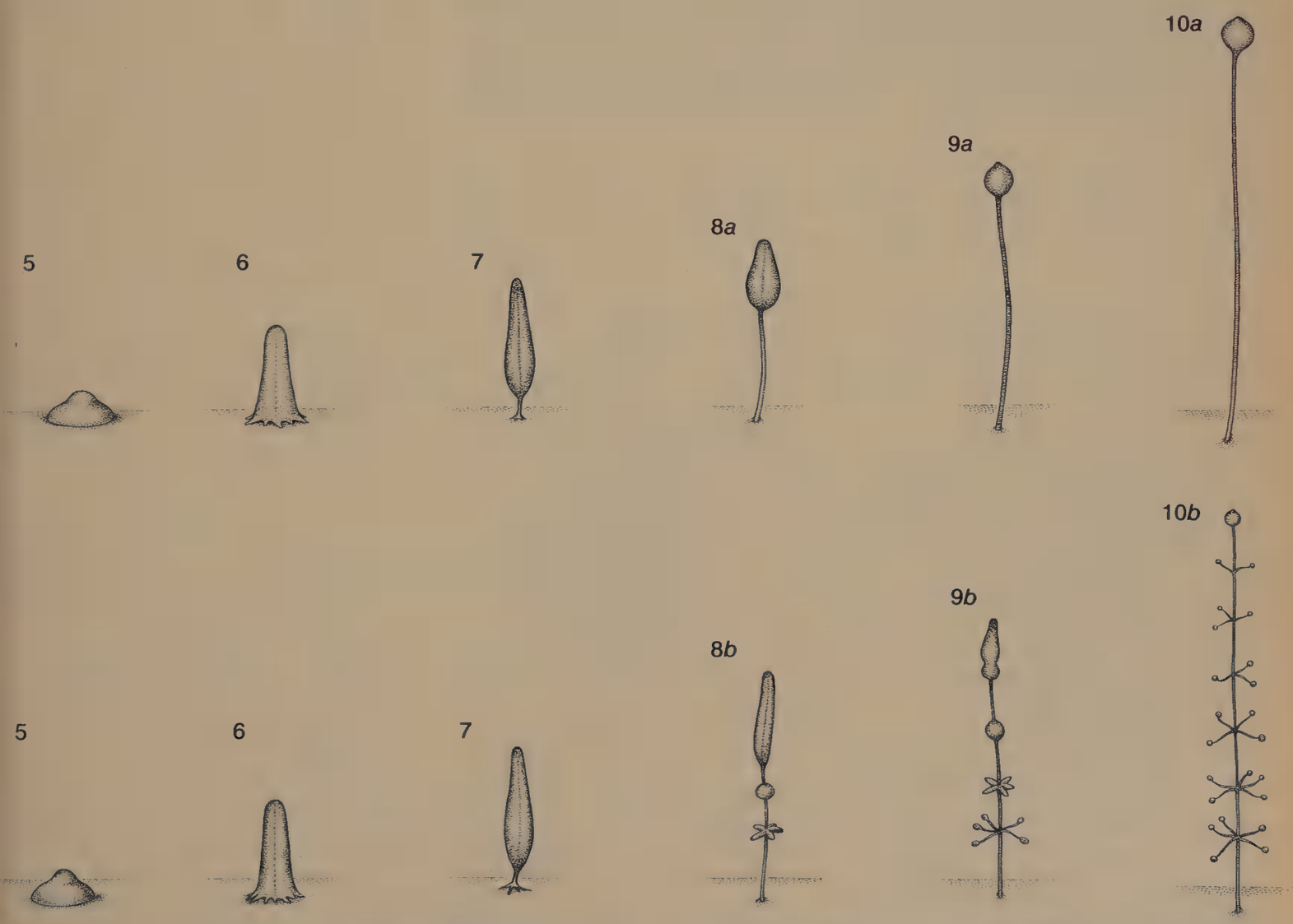
LIFE CYCLE AND MORPHOLOGY of two cellular slime molds, *Dictyostelium mucoroides* and *Polysphondylium violaceum,* are similar. They feed individual amoebae (*1*) until lack of food (bacteria) triggers either a group of cells (*2a*) or single cell (*2b*) to secrete chemical substance, called acrasin, that attracts surrounding amoebae (*3*). The cells stream toward a

The segregation phenomenon was noted in 1940 by Kenneth B. Raper and Charles Thom, who were then working in the U.S. Department of Agriculture. They grew cultures of the bacterium *Escherichia coli* and inoculated a culture dish with two species of social amoebae, *Dictyostelium discoideum* and *Polysphondylium violaceum*. The amoebae multiplied and grazed their way outward from the point where each species had been inoculated; eventually the expanding populations were thoroughly mixed. Having consumed the available bacteria, the amoebae in the region of mixing began to aggregate. As Raper and Thom wrote, "they regularly aggregated to different centers with the converging streams of the two forms commonly overlapping.... It was obvious that the stimuli in different species differed qualitatively." Raper and Thom recognized why that had to be. If the two species are together in the same soil, they must have a way to avoid coming together in the same aggregate in order to remain distinct species. Raper and Thom could not know how the two stimuli differed because the nature of the stimuli was not then known.

At the turn of the century scientists studying cellular slime molds had assumed that the amoebae must be brought together by some form of chemotaxis, but in the 1940's accounting for a developmental process by chemical attraction was viewed with suspicion; some form of "contact guidance" was a much more popular explanation. In 1942 Ernest H. Runyon of Agnes Scott College made it clear, however, that contact among cells could not bring about aggregation. In a series of experiments over the next few years I was able to show that the central mass had to be emitting some diffusible attracting agent. It could have been heat, but a number of indications made it clear that the agent must be a chemical substance. I called the substance an acrasin. (Acrasia is the name of a witch in Edmund Spenser's *Faerie Queene* who attracted men and transformed them into beasts, and *Dictyostelium* is a member of the Order Acrasiales.)

Over the years studies in numerous laboratories, including our own at Princeton University, have revealed many details of the acrasin system. Starvation is the stimulus for the switch from the feeding state to the social one. It sets in motion a series of chemical events within the individual cells, which soon begin to secrete an acrasin. At the same time the amoebae synthesize and display on their surface a large number of specific protein molecules that serve as receptors for the acrasin. The cells also secrete an acrasinase (an enzyme that inactivates the acrasin) and an inhibitor of the acrasinase. The acrasin, the acrasinase and the inhibitor combine to form a complex control system that modulates aggregation as it proceeds. To take just one example, the acrasinase keeps the external acrasin level generated by an individual amoeba low enough so that the cell can sense acrasin emitted by other cells.

Aggregation begins differently in different species, as Brian M. Shaffer, who was then working at the University of Cambridge, found some years ago. In



central collection point and pile up to form a multicellular organism (*4*). A tip forms on the cell mass (*5*) and the cells begin to differentiate. Some of them form a rigid internal stalk (*6, 7*). The remainder are borne aloft on the stalk to become spores. In *Dictyostelium* there is a single spherical mass of spores at the tip of a mature fruiting body (*8a, 9a, 10a*). In *Polysphondylium* there are whorls of small fruiting bodies in addition to an apical spore mass on the central stalk (*8b, 9b, 10b*). The spores become dispersed and split open, freeing amoebae.

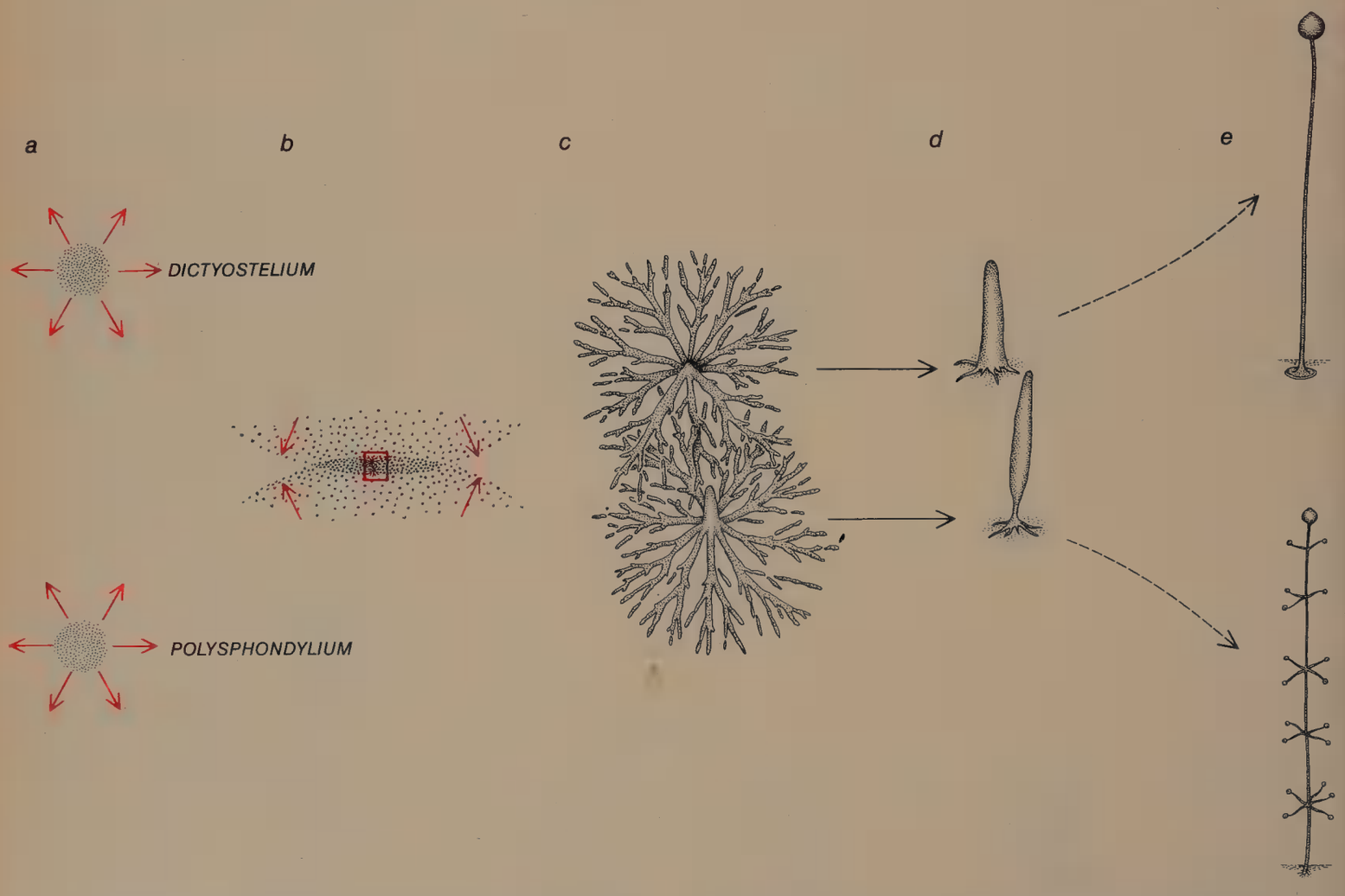
Polysphondylium a single cell in the population suddenly rounds up and begins to secrete acrasin. Neighboring cells streak in to join it, oriented by the concentration gradient of the secreted acrasin. In *Dictyostelium,* on the other hand, a small group of cells (Shaffer called it a cloud) seems to initiate aggregation. These cells first become relatively immobile and then clump together, and the clump attracts the surrounding cells. In both species the aggregating amoebae often come inward in pulses. There seems to be a good reason: to establish an effective overall gradient of acrasin that will orient distant cells, a good-sized central mass is required. The pulses can orient cells before such a gradient is established. If one amoeba gives off a puff of acrasin and the puff diffuses out to neighboring cells, it hits each cell on one side first, filling more of the cell's receptors on that side than on other sides, and thereby orients the cell. The responding amoeba in turn emits a puff that attracts cells beyond it.

When chemotaxis was firmly established as the mechanism of aggregation, I argued at first that it was not important to know the chemical nature of the acrasins. I say this with some embarrassment because the fact is that the moment the chemical structure of a key substance in physiology or development is discovered, one goes from the dark ages to modern times. Certainly that was true in the case of the *Dictyostelium* acrasin. In any event we soon joined other laboratories in an effort to characterize its attractant. It took from 1947 to 1967. I have told the story in detail before [see "Hormones in Social Amoebae and Mammals," by John Tyler Bonner; SCIENTIFIC AMERICAN, June, 1969], and I shall only summarize it here.

The major difficulty was developing a convenient bioassay: a way to test the biological effect of successively purified substances. Such assays were finally developed by Theo M. Konijn (who is now at the University of Leiden) and by us. We both did preliminary work toward identifying a factor found in bacteria that seemed to attract amoebae. Konijn came to Princeton, and soon after he arrived David S. Barkley, who was then a graduate student, had an inspiration: Why not try cyclic adenosine monophosphate (cyclic AMP), a nucleotide that Earl W. Sutherland, Jr., of the Vanderbilt University School of Medicine had discovered and had shown to act in mammalian cells as a "second messenger," mediating between hormones and cell functions.

It worked, with remarkable potency, in our chemotaxis assays. We went on to show that *Dictyostelium* cells not only respond to a minute amount of cyclic AMP but also produce it. (Now, to be sure, it is well known that almost all cells make cyclic AMP; not producing it is the rarity.) More to the point, we found that during aggregation the amoebae secrete the nucleotide in particularly large quantities (a hundredfold increase) and also are about 100 times more sensitive to the cyclic AMP than they are at other times.

The discovery that cyclic AMP is the natural acrasin of *Dictyostelium* opened up new lines of investigation. Workers in laboratories all over the world have studied just how the nucleotide orients cells and have analyzed the activity of adenyl cyclase, the enzyme that catalyzes the synthesis of cyclic AMP, and of phosphodiesterase, the acrasinase for this acrasin; they have examined the



TWO SLIME-MOLD SPECIES aggregate separately, ■■ Kenneth B. Raper and Charles Thom demonstrated in 1940 with this experiment. A culture of the bacterium *Escherichia coli* is inoculated with individual cells of *Dictyostelium discoideum* (*top*) and *P. violaceum* (*bottom*). The amoebae graze outward from the point of inoculation (*a*), become mixed and begin to aggregate (*b*). When the aggregations are examined under the microscope (*c*), they are seen to be separate. Streams of amoebae of the two species head for two different collection centers and form separate organisms (*d*), which develop to form distinctive fruiting bodies that are characteristic of the two species (*e*).

cell-surface receptors for cyclic AMP and the interactions of all the elements of the system. In 15 years there has been great progress toward understanding how some aspects of the development of *Dictyostelium* are controlled, and the new knowledge has implications not only for the development of the cellular slime molds but also for development in general.

As soon as we knew that cyclic AMP attracts *Dictyostelium* cells we tried it on *Polysphondylium*. It had no effect, just as Raper and Thom might have predicted (and as Shaffer later had predicted). The two species have different acrasins. The ensuing effort to identify the *Polysphondylium* acrasin was very different from our *Dictyostelium* experience. There was no clever guess by a bright graduate student. We could not even find a good source of unpurified attractant; bacterial products and even human urine had been good sources of cyclic AMP. We tried a variety of things that are generally rich in hormones and other active substances—milk, peptones (mixtures of peptides), yeast extracts and many others—with no effect. We had no recourse but to set about extracting the material from aggregating amoebae. The moral is that one cannot be blessed with fantastic luck every time.

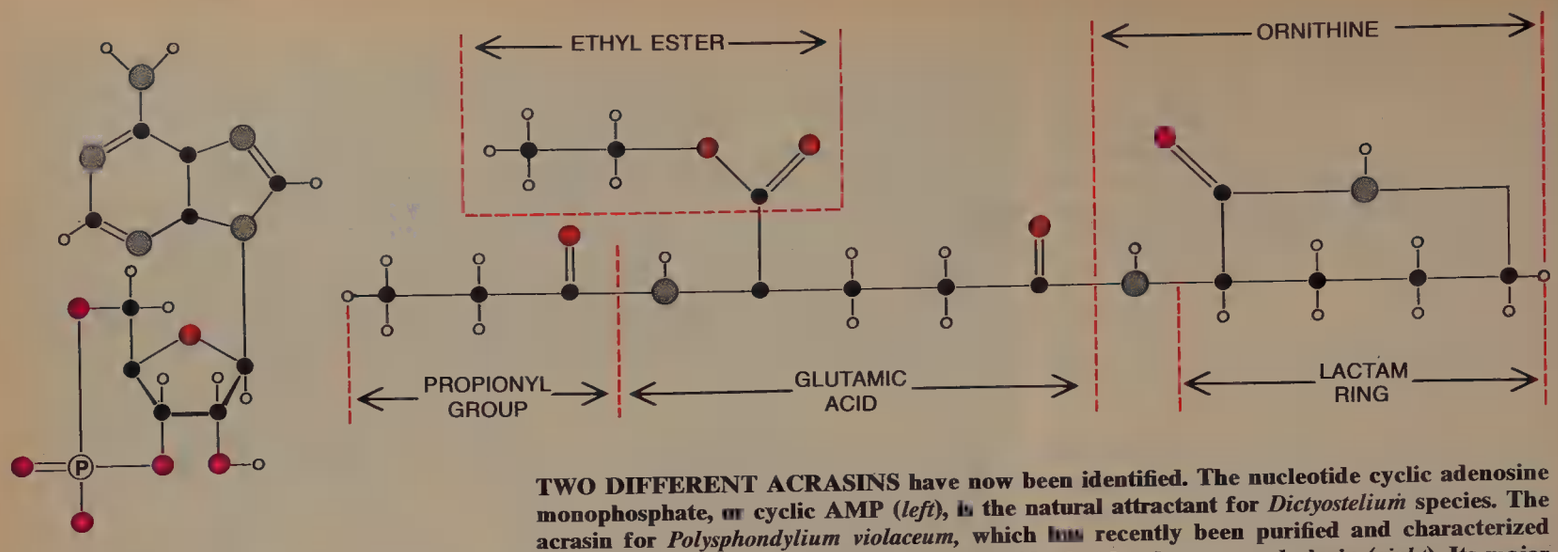
The first step was to find out how best to extract the acrasin from an aggregation of *Polysphondylium*. As we had learned working on *Dictyostelium*, the acrasinase secreted along with acrasin tends to break down the attractant as soon as it is secreted. We finally settled on a method devised by David W. Francis, who was then a postdoctoral fellow at Princeton and is now at the University of Delaware. It consists of dumping 40 percent ethyl alcohol directly on petri dishes containing cells in midaggregation, washing the dishes with more alcohol and then collecting the wash in a beaker. We boil off the alcohol and centrifuge off the precipitate that forms. A crude extract of acrasin is left in solution; it remains intact because the alcohol has denatured the acrasinase.

The remarkable thing is that the amoebae do not seem to be bothered by this 80-proof dunking; they go right on with their development after the alcohol has been removed. In order to collect as much crude extract of *Polysphondylium* acrasin as possible, over the years we have developed a routine that begins with 150 large petri dishes of aggregating amoebae, which cover every available surface in the laboratory. It has the disadvantage of stopping all other experiments for lack of space, but the laboratory does have a rather pleasant aroma. At first we used to think we could smell the acrasin, but it turned out to be the alcohol.

SEPARATE AGGREGATION of two species was recorded in this photomicrograph, made by Raper and Thom, in which the streams of aggregating amoebae are enlarged some 45 diameters. There is a small aggregation of *D. discoideum* (*top*) and a larger aggregation, with two dark centers, of *P. violaceum*. The inflowing streams of amoebae overlap, with the cells of each species apparently ignoring those of the other species as they head for their own center.

In the early days of this project, almost a decade ago, we sought the major characteristics of the *Polysphondylium* acrasin. This was done by Bernd Wurster (who is now at the University of Konstanz in Germany) and Pauline Pan (now with the Pennwalt Corporation). After more than two years of work we could say that the acrasin was a small molecule, with a molecular weight of less than 1,500 (estimated from its movement on a gel filtration column); that it was stable when heated; that it did not appear to be ionic in nature (that is, that the molecule was uncharged), and that it did not have any free amino or carboxyl groups, which are usually found respectively at the beginning and the end of a protein chain. Yet two enzymes—one protease and one peptidase—of a number we tried completely inactivated the acrasin, suggesting it was some kind of peptide (a short piece of a protein chain). Unfortunately this was not a conclusive finding because the enzyme preparations were impure; something quite secondary could have been playing a role. These were, however, neutral enzymes, which attack uncharged polypeptides, and this was consistent with the observation that the acrasin molecule seemed not to be charged. We concluded that the *Polysphondylium* acrasin might be a peptide in which there was no free amino group, with some evidence that it incorporated what is called an ester bond.

Further analysis was difficult because in spite of our best efforts our purest extracts were very impure. It was at this point that our colleague Osamu Shimomura, who is now at the Marine Biological Laboratory in Woods Hole, Mass., became interested and took charge of all the chemistry, including the purifications. To give him enough material to work with we had to increase the number of harvests. Under the direction of Hannah B. Suthers we started an "acrasin factory": two days of the week were devoted entirely to harvesting the 150 large petri dishes of acrasin. This went on over three years, for a total of 96 harvests. In the first two years we were still experimenting, perfecting the method of purification. In the third year we made what turned out to be the final push to accumulate crude acrasin, and

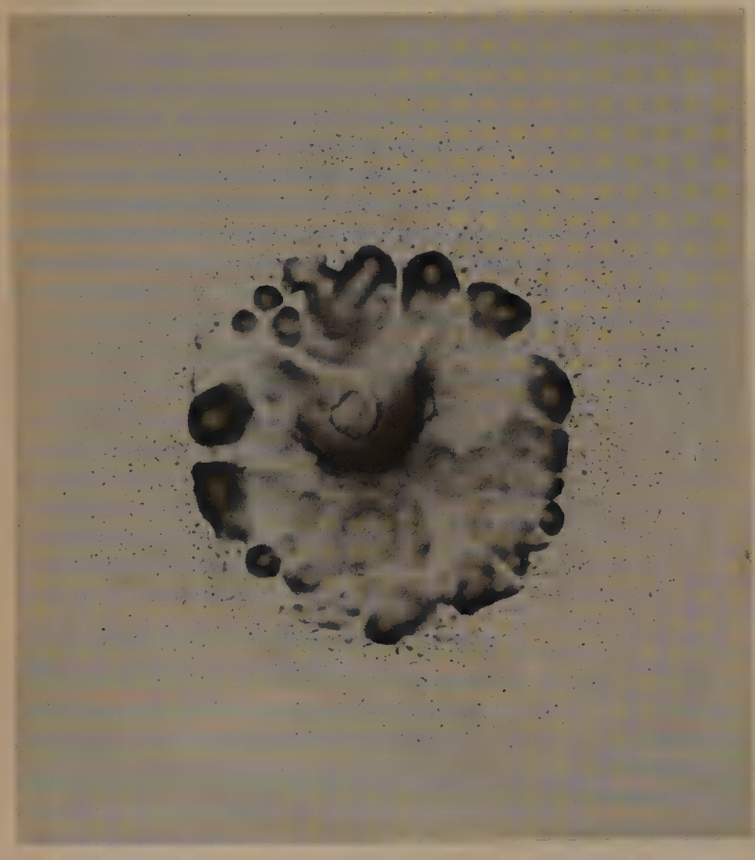


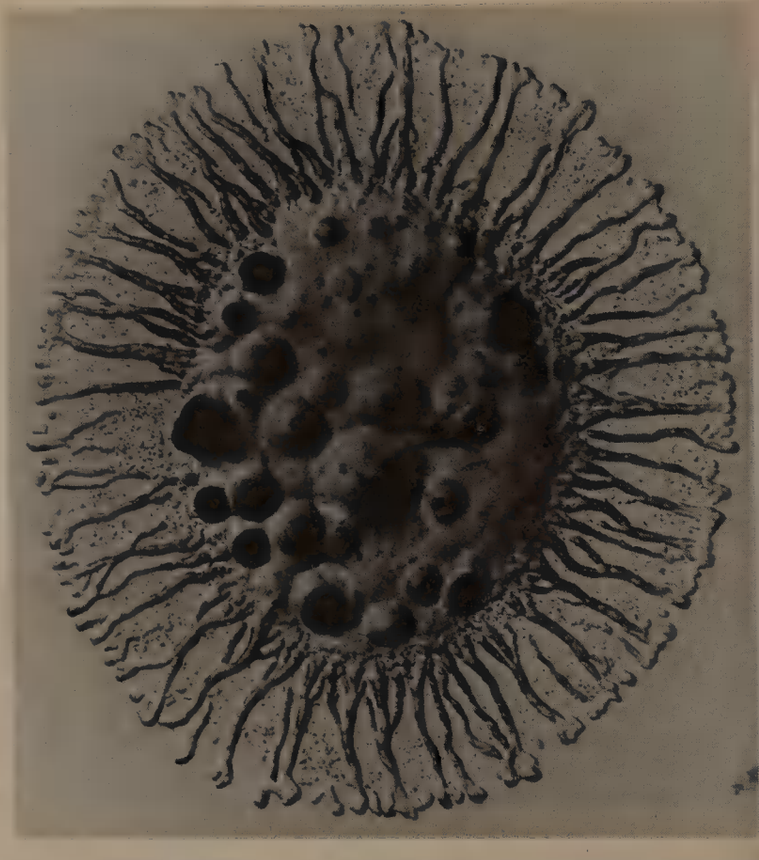
TWO DIFFERENT ACRASINS have now been identified. The nucleotide cyclic adenosine monophosphate, or cyclic AMP (*left*), is the natural attractant for *Dictyostelium* species. The acrasin for *Polysphondylium violaceum*, which has recently been purified and characterized by the author and his colleagues, is a dipeptide that has been named glorin (*right*). Its major components are the amino acids glutamic acid and ornithine. An amino group (NH_3) and a carboxyl group (COOH) of the glutamic acid are blocked respectively by a propionyl group and an ethyl ester. An amino group on the ornithine molecule is blocked by a lactam ring. Both cyclic AMP and glorin are small molecules, with molecular weights of 324 and 327 respectively.

40 collections gave us an initial four grams of very impure material. Shimomura put this through a series of nine purification steps involving three different kinds of gel filtration columns and numerous solvent systems. At the end we had 92 micrograms (millionths of a gram) of acrasin, about 98 percent pure.

The 92 micrograms of acrasin was divided into lots of about 10 micrograms that were sent away for different kinds of analysis. From the infrared spectrum we learned that there was indeed the strong possibility of an ester group. Amino acid analysis confirmed an earlier finding that the molecule contained two amino acids, glutamic acid and ornithine (which is not one of the usual components of proteins), in roughly equal amounts. The excitement came from the results obtained by Catherine E. Costello of the mass-spectrometry facility at the Massachusetts Institute of Technology. Using the latest magic (ionization techniques of field desorption and fast-atom bombardment, at both low and high resolution), it was possible in one final push to know the complete structural formula and the molecular weight. This is quite remarkable, considering that it comes from looking at a sample of about 10 micrograms of material.

The *Polysphondylium* acrasin turned out to be a dipeptide. There is a glutamic acid molecule with two side groups: a propionyl group, which blocks the amino end of the amino acid, and an eth-



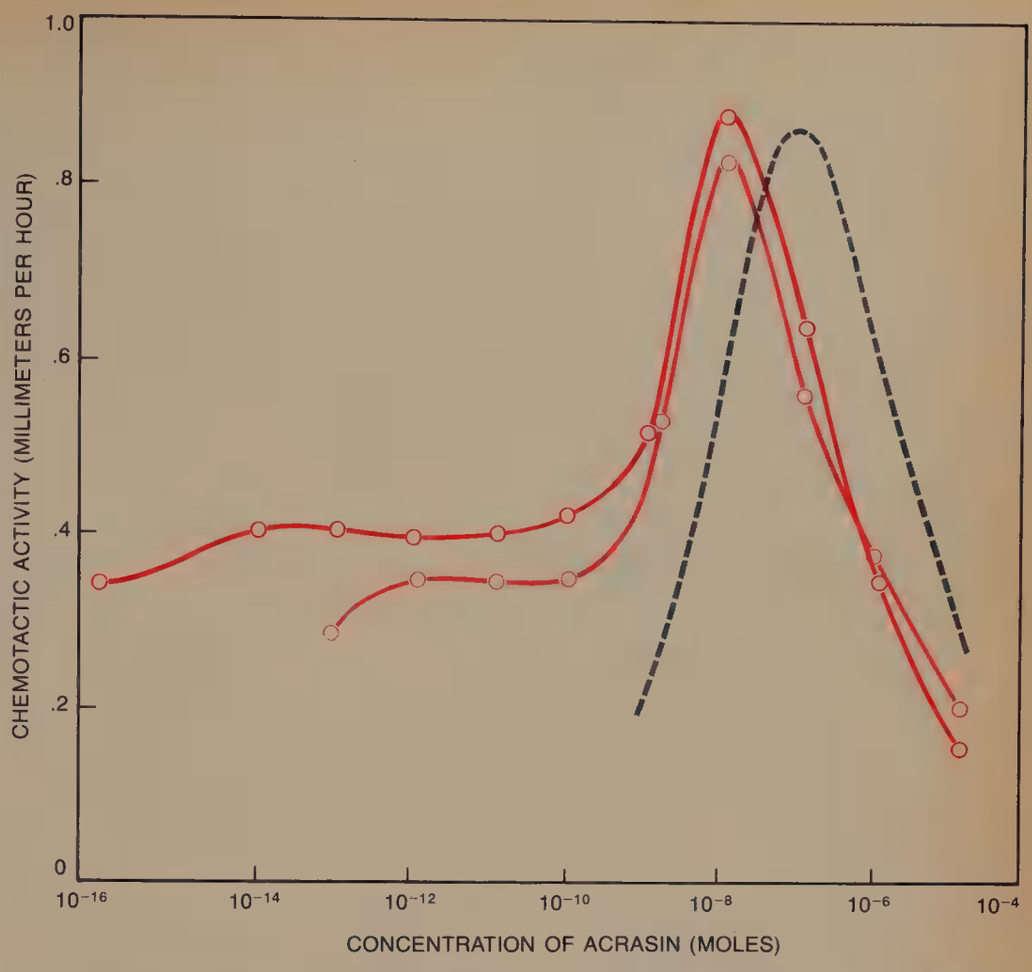


BIOASSAY demonstrates that glorin is indeed the natural attractant for *P. violaceum*. When a large clump of amoebae is placed on a block of agar (*left*), all but a few of the cells stay in the clump. When a clump of amoebae is placed on a block of agar that has been impregnated with synthetic glorin (*right*), the amoebae stream outward vigorously, attracted by the glorin. (The glorin directly under the clump is rapidly destroyed by acrasinase, an inactivating enzyme secreted by the cells. *P. violaceum*'s acrasinase has not yet been identified.)

yl ester, which blocks the free carboxyl group. The glutamic acid is attached by a peptide bond to an ornithine, but the ornithine is closed into a lactam ring, which again blocks a free amino group. The molecular weight of the acrasin is 327. Because it consists of glutamic acid and ornithine, whose abbreviations are *Glu* and *Orn*, we decided to call the acrasin glorin.

To confirm that glorin is the naturally occurring acrasin of *Polysphondylium* we needed to have the dipeptide synthesized and then test the manmade version in our chemotaxis assay with *Polysphondylium* amoebae. We sent our glorin formula to Peninsular Laboratories, Inc., and after a month or so they delivered to us 50 milligrams of synthetic glorin. (We calculated that it would have taken more than 500 years to accumulate that much material by extraction and purification in our laboratory!) At M.I.T., Costello immediately showed that its mass-spectral properties were identical with those of the natural, purified glorin. At Princeton, with considerable anxiety, we undertook the bioassay.

The result was unequivocal. The synthetic glorin had enormous biological activity: when it was placed in agar surrounding a mass of test amoebae, they streamed in a most dramatic way [*see bottom illustration on opposite page*]. In a quantitative test we compared the synthetic glorin's ability to attract amoebae with that of the natural, purified glorin (the remains of our precious hoard) and found the two were essentially identical in their activity. When we compared the glorin curves with data showing the effect of cyclic AMP on aggregating *Dictyostelium* amoebae, it was clear that glorin works effectively at even more dilute concentrations than cyclic AMP does.



CHEMOTACTIC ACTIVITY of glorin synthesized in the laboratory (*dark color*) is compared with the activity of the natural attractant (*light color*) purified from aggregating cells. The two curves are essentially identical, confirming the chemical analysis of the natural glorin. A similar curve (*gray*) shows the effect of cyclic AMP on aggregating cells of *D. discoideum*.

So much for the excitement of the chase. What can we now do with glorin? One thing is to work out all the details of the glorin aggregation system, including the identification of its receptors, the acrasinase and the inhibitor of the acrasinase. We also want to ask this question: What are the respective contributions of glorin and of cyclic AMP to differentiation? The reason for this interest is that is has been known since our early experiments with Konijn that cyclic AMP is present in *Polysphondylium*. Later, in our laboratory and in others, mounting evidence was found that cyclic AMP is involved in the differentiation of both stalk cells and spores in *Dictyostelium*, and there is similar evidence in *Polysphondylium*. There is even interesting work in *Dictyostelium* showing that cyclic AMP is directly involved in controlling the expression of some developmental genes. Recent work in laboratories in the Netherlands and the U.S. has provided convincing evidence that chemotaxis continues to operate in the cell mass during later stages of development: a cyclic AMP gradient within the multicellular organism orients the cells, thereby guiding the direction of the advancing tip. During these later stages of *Dictyostelium* development, then, cyclic AMP is involved in two things, chemotaxis and differentiation, at the same time and in the same place.

We want to find out if in *Polysphondylium* there is perhaps a separation of functions, with chemotaxis controlled by glorin and differentiation controlled by cyclic AMP. The relation of the two signaling systems may not be simple and separate; it is quite possible that they are mutually dependent on one another or that glorin is more directly involved in cell differentiation than we currently assume. Here lies a unique opportunity to compare two parallel developmental systems that have different chemical signaling systems. Among other things, it should provide a way to separate, isolate and identify the controls of the different developmental events.

Perhaps the main point is that *Dictyostelium* and *Polysphondylium* live side by side in the soil, are similar in appearance and go through remarkably similar stages of development. Yet they have totally different signaling systems, one system based on a nucleotide and the other on a peptide. The only comparable situation is found in the body of animals, including human beings, where cyclic AMP acts in many signaling reactions, both hormonal and neural, and a number of peptides function in the nervous system as neurotransmitters. The presence of rival signal systems based on small nucleotides and on peptides is a very general biological phenomenon, not just a slime-mold curiosity.

Some broad evolutionary questions arise from the fact that these two signaling systems are so different. I pointed out above that Raper and Thom's experiments seemed to imply that by responding to two different stimuli two species living in the same place can keep from intermingling and thereby losing their identity. The problem extends beyond two species. James C. Cavender of Ohio University and Raper, who have spent much time collecting slime molds all over the world, have shown that in some temperate and tropical environments as many as from five to eight species may coexist. It is not necessary for all of them to have different acrasins to keep separate. As Raper and Thom showed in their original paper, there can also be "cell-surface incompatibility": some

Discwasher Wants to Make You a Video Authority

For your copy of "THE VIDEO REFERENCE", complete the coupon below and return it to Discwasher, along with a check or money order for $1 to cover postage and handling.

Please send me _____ copies of "THE VIDEO REFERENCE". Enclosed is my check or money order for $1/copy to cover postage and handling.

Name

Address

City State Zip

discwasher

1407 NORTH PROVIDENCE ROAD, P.O. BOX 6021, DEPT. SA, COLUMBIA, MO 65205 USA

A DIVISION OF JENSEN an ESMARK Company

STALKS AND SPORE MASSES of *Polysphondylium pallidum* (which, like *P. violaceum*, secretes and responds to the acrasin glorin) are enlarged 16 diameters in a photomicrograph made by Raper. They have developed at the center of a bacterial colony, whose edge is seen at the top of the picture. The fruiting bodies are in different stages of development: the larger ones are still building and the one at the right, with a spherical apical spore mass, is mature.

species aggregate together, in response to the same acrasin, and then the cells of the two species separate from each other at the center of the aggregate to form two separate multicellular individuals. Nevertheless, there is already evidence that there are at least eight different acrasins among the approximately 50 known species of cellular slime molds; when all these species have been carefully investigated, additional acrasins will surely be revealed.

This means that the extent to which cellular slime molds have been able to evolve from one chemotaxis system to another is quite astounding. It raises the fascinating question of how a signaling system can shift from one such chemical system, with its acrasin, its receptors for a specific acrasin and no doubt other associated substances, to another system. Clearly the selection pressure on slime molds to diversify and occupy new niches must be intense; even given that assumption, how are the evolutionary steps of change from a cyclic AMP chemotaxis system to a glorin chemotaxis system (or vice versa) accomplished?

I find it useful to compare slime molds in the soil to signal systems in man. Human beings have a vast array of hormones and neurotransmitters, and one assumes they arose during man's evolution from simpler animals that had far fewer signal systems. This leads to the same question as before: How did new hormones and neurotransmitters arise during the early evolution of animals? Again one could imagine that for increased control it is important to have a series of quite separate hormones and neurotransmitters, and that through natural selection some of those signaling systems gave rise to new ones, just as slime molds evolved into new species.

I am drawing a grand analogy between a mammalian body and a plot of soil and postulating for both a trend toward the increased diversification of chemical signals. It is therefore conceivable that the origins of acrasins may help us to understand the origins of animal neural and hormonal signaling systems. There is, however, a difference of staggering proportion between slime molds and mammals: the signals in animals, including human beings, produce an individual organism with extraordinarily unified accomplishments, including noble thoughts, something one can hardly expect of a shovelful of soil.

Does Rain Mean Wet?

Not with the Gore-Tex™ Rain Hat from REI. The unique qualities of Gore-Tex™ allow fabric to breathe and air to circulate, so your head stays dry and comfortable. We guarantee satisfaction or your money back! **Offer expires August 15, 1983.**
Sizes: S, M, L, XL. **$10.95**ppd

Please send me _____ Gore-Tex™ Rain Hat(s). Size(s): _____
Order #C40-291 (Check color):
☐ Royal(**2**) ☐ Navy(**4**)
☐ Forest Green(**5**) ☐ Slate Blue(**6**)
☐ Tan(**8**) ☐ Chili Bean(**9**)
☐ Send me a free color catalog.

Name _____
Address _____
City _____
State _____ **ZIP** _____
☐ Check ☐ Mastercard ☐ Visa
Acct # _____
Exp. Date _____
Call toll free: **1-800-426-4840.**
In WA state call: **1-800-562-4894.**

REI **P.O. Box C-88127 Seattle, WA [illegible]**
Quality Outdoor Gear and Clothing Since 1938

Where can you get a front row seat to hear the National Symphony every week?

The Mutual Broadcasting System now brings the National Symphony as close as your radio. Under the musical direction of Mstislav Rostropovich, you'll hear a brilliant concert season, arranged just for you.

Once a week, on your Mutual radio station, you'll hear the classics you love—Mozart, Brahms, Beethoven and more. You'll hear performances from the Kennedy Center for the Performing Arts in Washington, D.C., and concerts from the Symphony's recent European Tour. Some of these concerts will even be coming to you live, via satellite, in stereo.

Check your local radio listings for the next performance. Then take the best seat in your house. Because for the best in classical music, the answer is Mutual, your full-service network.

The answer is Mutual.

MUTUAL BROADCASTING SYSTEM

Intuitive Physics

Although Newton's laws are well known, tests show many people believe moving objects behave otherwise. The subjects of the tests tend to follow a theory held in the three centuries before Newton

by Michael McCloskey

Imagine the following situation. A person is holding a stone at shoulder height while walking forward at a brisk pace. What will happen if the person drops the stone? What kind of path will the stone follow as it falls?

Many people to whom this problem is presented answer that the stone will fall straight down, striking the ground directly under the point where it was dropped. A few are even convinced that the falling stone will travel backward and land behind the point of its release. In reality the stone will move forward as it falls, landing a few feet ahead of the release point. Newtonian mechanics explains that when the stone is dropped, it continues to move forward at the same speed as the walking person, because (ignoring air resistance) no force is acting to change its horizontal velocity. As the stone travels forward it also moves down at a steadily increasing speed. The forward and downward motions combine in a path that closely approximates a parabola.

The discrepancy between intuition and physical reality in this situation is somewhat surprising. One might expect that as a result of everyday experience people would have reasonably accurate ideas about the motion of objects in familiar situations. One might even suppose people would have an informal grasp of the general principles that govern objects in motion. It seems that such is not the case. Recent studies on the nature, development and application of knowledge about motion indicate that many people have striking misconceptions about the motion of objects in apparently simple circumstances. The misconceptions appear to be grounded in a systematic, intuitive theory of motion that is inconsistent with fundamental principles of Newtonian mechanics. Curiously, the intuitive theory resembles a theory of mechanics that was widely held by philosophers in the three centuries before Newton.

In a series of studies done by my colleagues and me at Johns Hopkins University, high school and college students were asked to predict how a moving object would behave in diverse situations. One experiment concerned the motion of an object that is initially confined to a circular path and then is suddenly released. Such motion takes place, for example, when an athlete throws a discus. Alfonso Caramazza, Bert F. Green, Jr., and I asked college students to consider a situation in which a man ties a ball to the end of a string and whirls them in a circle above his head. What would happen to the ball, we asked, if the string were to break? We gave the students a diagram showing the circular path of the ball, and we asked them to draw the path it would follow after the string broke, ignoring air resistance. We told them to show only the horizontal motion of the ball and not to try to depict the ball's downward motion.

It is a fundamental principle of Newtonian mechanics that an object moves in a straight line unless it is acted on by an external force. Therefore the correct answer to the problem is that the ball travels in a straight line along a tangent to its original circular path, beginning at the moment the string breaks. Roughly a third of the 50 students to whom we presented the problem, however, drew a curved trajectory for the ball. That is, they believed the curvilinear motion of the ball would persist for some time even after the ball was released from the constraint imposed by the string.

Experiences reported by some of our subjects soon led us to suspect that intuitive beliefs about motion play a role not only in people's thinking about hypothetical situations but also in their interactions with real objects. For example, a subject in one of our studies told us that the first time he threw a stone with a sling he broke a window because after the stone was released it did not curve as he expected it to. In order to investigate the relation between beliefs and actions systematically, my student Deborah Kohl and I devised a simple situation in which intuitive ideas about circular motion might influence how people performed a task.

We asked a group of subjects to push a small object across a table in such a way that it would pass through a 90-degree segment of a circular ring painted on the table. The object was a small "puck" with a ball bearing that allowed it to roll smoothly across the tabletop. The ring segment was bounded by two concentric circular arcs and by two line segments that would meet, if they were extended, at the common center of the arcs. We told our subjects to start with the puck near the edge of the table. They were to push it up to one of the straight lines on the boundary of the ring segment and release it there; the puck was then to roll across to the other straight edge of the ring segment without touching either one of the arcs.

Many of our subjects realized that the puck could pass through the ring segment in a straight line, and so they tried to push the puck along a straight trajectory. A fourth of the subjects, however, moved the puck in a curved path before releasing it, in the belief that it would continue to curve after release and so would follow the arc of the ring segment. Most of the people who tried the second strategy were visibly surprised when the puck failed to curve.

It was relatively easy to construct a task in which intuitive beliefs about the motion of dropped objects might influence a person's actions. My student Allyson Washburn and I asked undergraduates to walk across a room and while they were walking to drop a golf ball so that it would hit a target marked on the floor. Almost half of the subjects tried to release the ball when it was directly over the target, in the belief that it would fall straight down. One subject deliberately walked past the target before dropping the ball, thinking the ball would move backward as it fell.

The failure to understand the role a carried object's initial motion plays in determining its path when it is dropped

or thrown may also lead to difficulties with other actions. Inexperienced basketball players systematically shoot the ball much too hard when they drive toward the basket, often slamming the ball against the backboard. Novice football quarterbacks often find it hard to throw a pass accurately while they are running. A recent story in the Baltimore *Sun* described the attempts of a man who could juggle while he was riding a unicycle to teach the skill to another unicyclist. The juggler told the trainee to throw the balls straight up as she moved forward on the unicycle, but the trainee questioned the appropriateness of the instruction on the grounds that the balls would fall behind her as the unicycle moved forward.

Why do people so often misjudge the path of a moving object when they solve problems or carry out actions? Several recent studies, including the ones I have described, indicate that the errors are not random but systematic. They arise from a general, coherent theory of motion that adequately guides action in many circumstances but is nonetheless at variance with the principles of Newtonian mechanics. It is therefore the misconceptions embodied in an intuitive physical theory that occasionally give rise to errors in judgment about motion. The intuitive theory bears a striking resemblance to the pre-Newtonian theory of impetus.

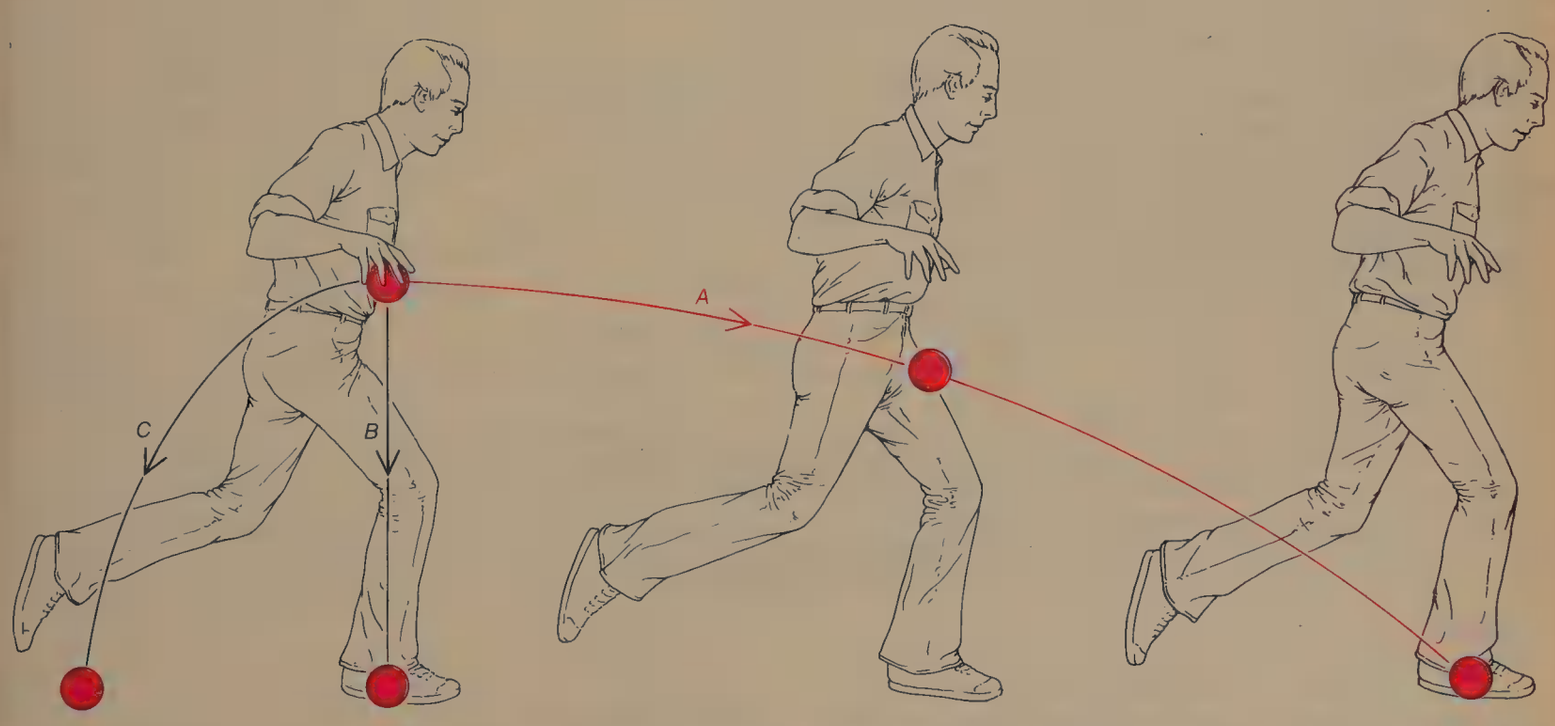
The impetus theory was a medieval correction to Aristotle's account of motion; it was discussed by the Greek scholar John Philoponus in the sixth century and was developed most fully by the French philosopher Jean Buridan and others in the 14th century. A succinct statement of the theory was given by Buridan (in translation): "When a mover sets a body in motion, he implants into it a certain impetus, that is, a certain force enabling the body to move in the direction in which the mover starts it, be it upward, downward, sideward or in a circle. It is because of this impetus that a stone moves on after the thrower has ceased moving it."

The medieval impetus theory shared with Aristotelian physics the idea that motion must have a cause. The continuous action of a force was thought to be necessary to keep an object in motion, a concept Newtonian mechanics explicitly denies. In Aristotelian physics the force responsible for the motion was necessarily external to the moving object. This requirement made it quite difficult to explain the motion of a projectile, because once a projectile leaves its launcher it is hard to find any source of external force. In the impetus theory the difficulty was circumvented by assuming that the motion of a projectile is maintained by a force internal to the object, which is acquired when the object is set in motion. The internal force was called impetus.

To account for the observation that most moving objects eventually come to a stop impetus theorists also assumed that impetus gradually dissipates. For example, according to the impetus theory, a ball rolling across a floor gradually comes to rest because the impetus acquired when it was set in motion steadily diminishes until it is entirely exhausted. Some impetus theorists held that impetus dissipates spontaneously, whereas others, including Buridan, argued that external influences such as air resistance are responsible for its loss.

The dissipation of impetus was an important consideration for the medieval analysis of projectile motion. The 11th-century Arab philosopher Avicenna held that a projectile travels in a straight line in the direction in which it is launched until the impetus from the launcher is entirely spent. The projectile then momentarily comes to rest and its natural heaviness supplies a downward impetus, causing it to fall straight to the ground. Hence for an object launched horizontally the trajectory is shaped like an inverted *L*: the object travels straight out, stops and then falls straight down.

A theory of projectile motion that gives a somewhat more accurate description of the trajectory was proposed in the 14th century by the scholastic philosopher Albert of Saxony. In Albert's account projectile motion has three phases. First the impetus of the projectile overpowers its natural heaviness, and the projectile moves along a straight line in the direction in which it was aimed. The impetus gradually dissipates, however, and at some point the projectile's weight causes it to begin to deviate downward from its original path. Thus in the second phase of its trajectory the projectile moves under



BALL DROPPED by a running person continues to move forward at the same speed as the runner. The forward motion combines with a steadily accelerating downward motion to produce a parabolic trajectory (*A*). Intuitive beliefs about the motion of objects do not always correspond with physical reality. The author and his colleagues asked college students where a ball would land if it were dropped by a walking person. Only 45 percent of the students knew the ball would travel forward as it fell. Forty-nine percent thought the ball would fall straight down and land directly under the point where it was released (*B*); 6 percent thought the ball would move backward as it fell (*C*).

the influence of both the original impetus and the downward impetus imparted by its weight. In the third phase the original impetus is entirely spent and the projectile falls straight down.

For the motion of a projectile thrown upward a similar account was given in the second century B.C. by the Greek astronomer Hipparchus. According to him, the upward impetus of the projectile initially predominates over its weight, and so the projectile moves upward; as the impetus diminishes, the upward motion slows. Eventually the upward impetus gets so weak that the weight of the projectile overcomes the impetus and the projectile begins to fall. As the upward impetus continues to dissipate, the projectile moves downward with increasing speed. In his early writings Galileo adopted Hipparchus' theory in order to discuss the acceleration of a falling object.

It is important to note that the medieval impetus theorists did not confine their concept of impetus to motion in a straight line. According to the impetus theory, an object that is moved in a circular path acquires circular impetus, which acts to maintain the circular motion. The concept of circular impetus was invoked to explain phenomena such as the continued rotation of a wheel after it is set in motion and the presumed rotation of the celestial spheres in which the stars and planets were thought to be fixed. Leonardo da Vinci applied the idea of circular impetus in order to describe the motion of an object under conditions quite similar to the ones we presented in the ball-and-string problem: "Everything movable thrown with fury through the air continues the motion of its mover; if, therefore, the latter move it in a circle and release it in the course of this motion, its movement will be curved."

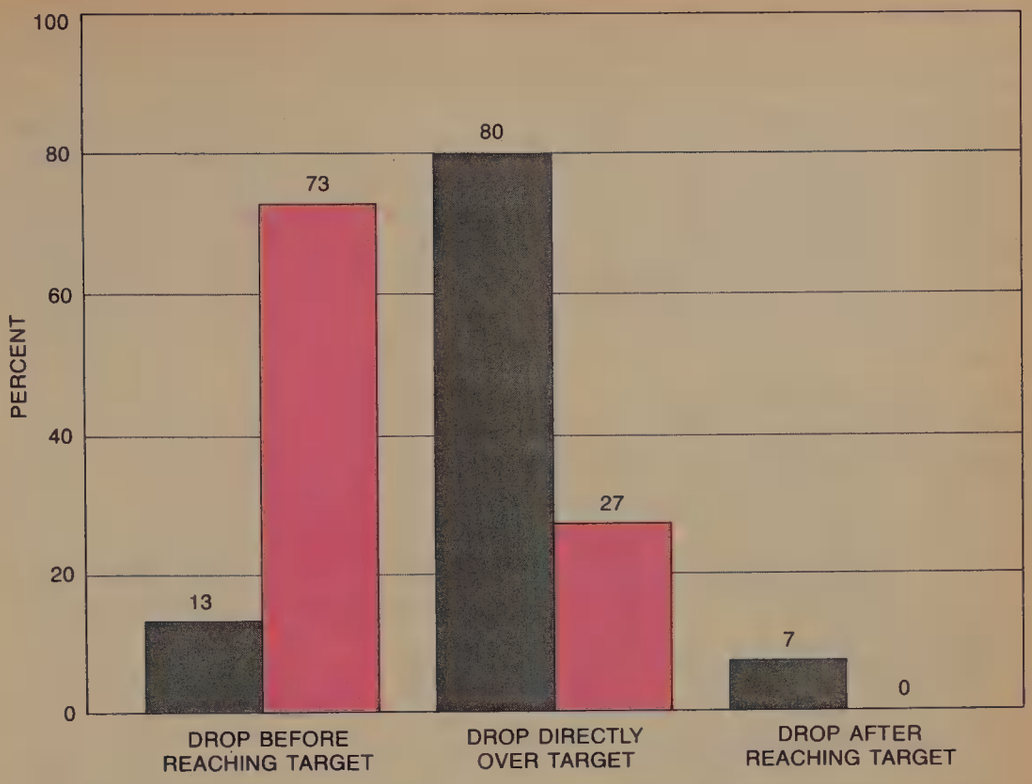
The medieval impetus theory is incompatible with Newtonian mechanics in several fundamental ways. Newton's first law of motion holds that just as no force is required to keep an object at rest, so no force is required to

IMPETUS THEORY OF MOTION, widely endorsed in the 14th through the 16th centuries, held that when an object is set in motion, it acquires an internal force called impetus that acts to keep it moving. The theory is contrary to fundamental principles of Newtonian mechanics, which hold that no force is needed to keep an object moving in a straight line at a constant speed. Nevertheless, many people still have an intuitive understanding of the motion of objects that is quite similar to the impetus theory. This drawing, which is reproduced from the 1582 edition of *Bawkunst Oder Architectur aller fürnemsten,* by Walther Hermann Ryff, illustrates a version of the impetus theory that divides the trajectory of a projectile into three phases. In the first phase the impetus provided by the cannon overpowers the weight of the projectile, and the projectile moves in a straight line. The impetus gradually dissipates, however, and its weight causes the projectile to begin to curve downward. In the second phase the projectile moves under the influence of the impetus from the cannon and the downward impetus imparted by its weight. In the third phase the impetus from the cannon is exhausted and the projectile falls straight down.

keep an object moving at a constant velocity (that is, at a constant speed in a straight line). In fact, in Newtonian mechanics a state of rest and a state of uniform velocity are regarded as being equivalent. No physical distinction can be drawn between them, and so any object that is not changing its speed or its direction can be described as being at rest or in uniform motion, depending on the choice of a frame of reference. For example, it is equally valid to describe a passenger in a nonaccelerating airplane as being at rest in the reference frame of the airplane or as being in motion in the reference frame of the earth. Coffee poured inside the airplane falls into a cup just as it does when it is poured at rest with respect to the ground; it does not splash back toward the rear of the airplane, as it would if the impetus theory were correct.

From the perspective of the impetus theory a fundamental distinction must be made between a state of rest and a state of motion: moving objects can have impetus but objects at rest cannot. As I have pointed out, in Newtonian mechanics no such distinction can be drawn. Moving objects do have momentum, however, and so in some superficial respects the concept of impetus resembles the concept of momentum. There are two crucial differences: the first is that impetus was considered to be a cause of motion, whereas momentum is merely a quantity employed to describe motion. The second is that in the impetus theory an object is viewed as having impetus in an absolute sense, whereas in Newtonian mechanics momentum, like velocity, is defined relative to a frame of reference.

It is also incorrect to suppose a rolling object stops or a projectile falls because of the loss of impetus. In the Newtonian framework moving objects come to a stop or begin to fall because external forces act to change the speed or direction of their motion. A ball rolling across a floor is slowed by friction, a force that acts in a direction opposite to the direction of the ball's motion. A projectile fired horizontally is accelerated downward by the constant force of the earth's gravity. The projectile's horizontal motion is independent of its vertical motion, although both are slowed by the resistance of the air. In a vacuum the horizontal velocity would remain constant from the moment the projectile was launched until the moment it hit the ground. For a projectile aimed above the horizontal the continuous action of gravity steadily reduces the initially positive upward velocity to zero and then reduces it to progressively larger negative values; in other words, after the projectile reaches its peak it is accelerated downward.



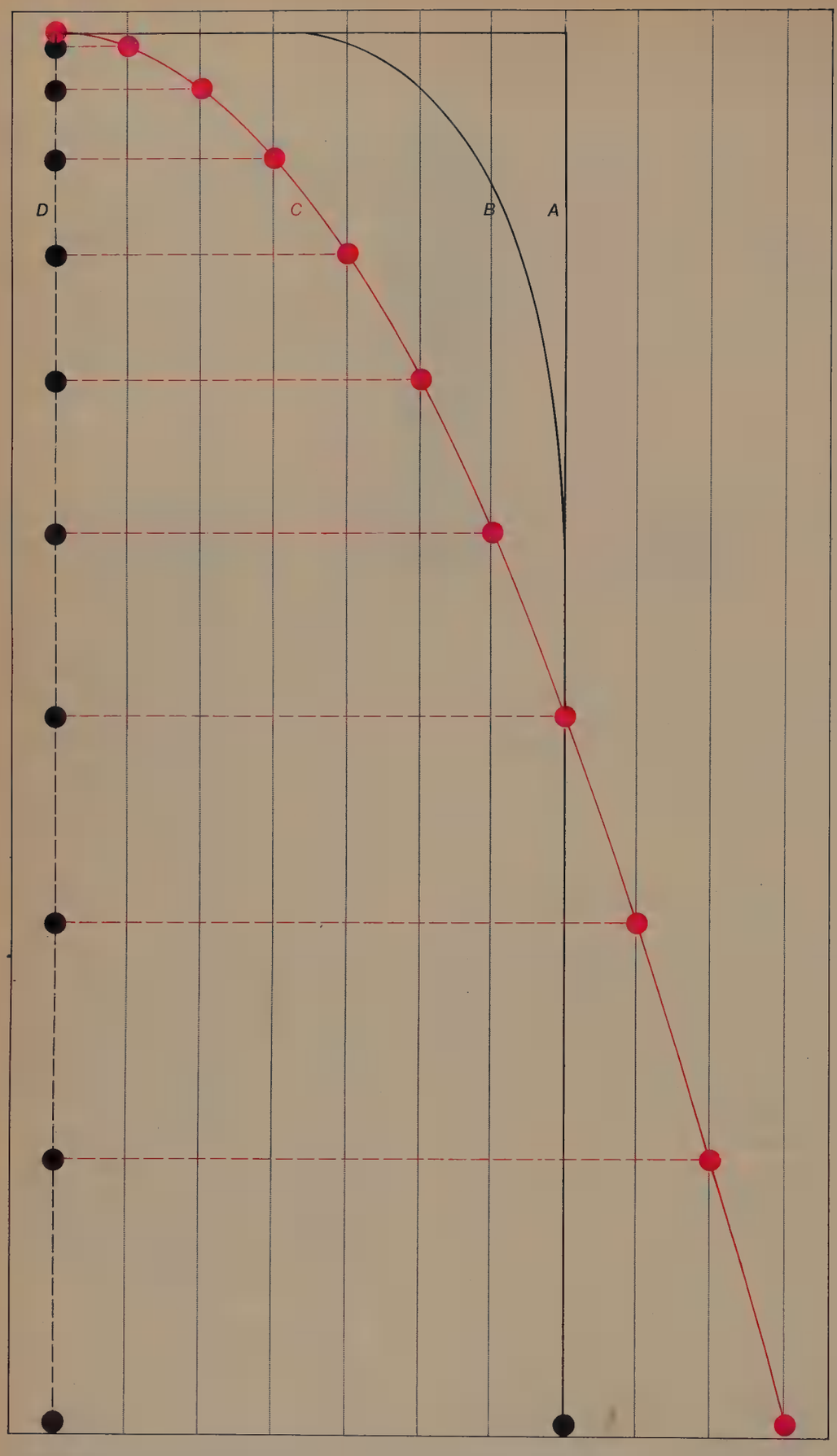
INTUITIVE BELIEFS about motion can affect the ways in which people interact with moving objects. College students were asked to walk across a room and while they were walking to drop a golf ball onto a target marked on the floor. The results of the test, summarized by the bar graphs, show that training in physics affects the strategies people adopt for hitting the target. Subjects who had taken at least one physics course (*bar graphs in color*) were likelier than those with no physics training (*bar graphs in black*) to release the ball before reaching the target. More than a fourth of the subjects trained in physics, however, incorrectly thought the ball should be dropped when it was directly over the target. The finding shows that students may often retain the intuitive impetus theory as a framework for interpreting course material on Newtonian mechanics and therefore distort the new material to fit the intuitive theory.

The effect of gravity on a projectile's vertical motion is not delayed by any horizontal component of motion; the projectile begins to deviate downward from the direction in which it was aimed the instant it is fired. Thus the impetus theorists were also incorrect in maintaining that a projectile travels along a straight line for some time after it is launched.

Finally, Newtonian mechanics and the impetus theory differ in their treatment of circular motion. In the impetus theory circular motion is viewed as being not fundamentally different from motion in a straight line. Both forms of motion are generated by imparting the appropriate impetus to an object. In Newtonian mechanics there is clear distinction between motion in a straight line and motion in a circle. An object moves uniformly in a straight line when no force is acting on it. For an object to move in a circle, however, it must be acted on constantly by an outside force that tends to deflect the motion from linearity. A ball whirled in a circle at the end of a string is acted on by the tension in the string, which continuously pulls on the ball and so keeps it from moving off at a constant speed in a straight line tangent to the circle.

Each major aspect of the impetus theory I have discussed has its counterpart among the beliefs about motion that are widely held today. Indeed, the ideas about motion held by most people with no formal training in physics, and by many who have completed at least one physics course, are much closer to the account given by the impetus theory than they are to Newtonian mechanics. These conclusions are supported by our studies at Johns Hopkins and by experiments carried out by John J. Clement of the University of Massachusetts at Amherst, by Audrey B. Champagne and Leopold E. Klopfer of the University of Pittsburgh and by Laurence Viennot of the University of Paris.

In one study, carried out by Kohl, Washburn and me, high school and college students tried to solve several problems about motion and then explained their answers. In many instances the students stated the impetus theory quite explicitly. For example, one student, who used the term momentum in explaining his answers, was asked to define the term. He responded: "It's something that carries [an object] along after a force on it has stopped. Let's call it the force of motion. It's something that keeps a body moving." Another student described the momentum of a ball: "[It is] a force that has been exerted and put into the ball so this ball



BELIEFS ABOUT PROJECTILE MOTION held by experimental subjects resemble those of medieval impetus theorists. Subjects were asked to draw the trajectory of a projectile launched horizontally, ignoring the resistance of the air. Five percent indicated that the projectile moves straight out and then straight down (*A*). The 11th-century Arab philosopher Avicenna held a similar view. Thirty-five percent of the subjects believed the projectile first travels in a straight horizontal line, then begins to curve downward and finally falls straight down (*B*). The description agrees with the three-phase theory of projectile motion proposed by the 14th-century scholastic philosopher Albert of Saxony. Twenty-eight percent of the subjects gave the correct trajectory, which is parabolic (*C*); most of the remaining 32 percent correctly indicated that the projectile begins to fall immediately but incorrectly thought it eventually falls straight down. The parabolic path represents the combination of a constant-velocity horizontal motion and accelerating vertical motion. The distance the projectile falls during each interval of time is the same as that of an object dropped from rest (*D*). Thus a bullet fired horizontally and a bullet that is dropped simultaneously from the same height hit ground at the same time.

now that it's traveling has a certain amount of force. The moving object has the force of momentum and since there's no force to oppose that force, it will continue on until it is opposed by something." In the two statements the belief that motion is maintained by an internal force is quite clear.

Also expressed by our subjects was the idea that moving objects slow down and stop because of the dissipation of impetus. One subject, asked to explain why a ball rolling along the floor would eventually come to a stop, gave the following explanation: "I understand that friction and air resistance adversely affect the speed of the ball, but not how. Whether they sort of absorb some of the force that's in the ball—I'm not sure. In other words, for the ball to plow through the air resistance or the friction, if it has to sort of expend force and therefore lose it, I'm not sure." The speaker clearly, although tentatively, assumes that friction and air resistance dissipate the impetus of the ball.

The intuitive idea that an object released from circular motion follows a curved path for some time thereafter appears to stem from a belief that such an object has circular impetus. A subject who drew a curved trajectory for our problem involving a ball whirled at the end of a string explained that the ball would curve "because of the directional momentum. You've got a force going around, and [after the string breaks, the ball] will follow the curve you've set it on until the ball runs out of the force within it that you've created by swinging." In other words, the whirled ball acquires circular impetus, which causes it to follow a curved path for some time after the string breaks. The circular impetus dissipates, however, and so the trajectory of the ball gradually straightens out.

Our subjects also recapitulated the medieval impetus theory in their beliefs about projectile motion. When subjects were asked to explain why projectiles rise to a peak and then fall, they frequently echoed the account initially proposed by Hipparchus. For example, one subject explained the vertical motion of a cannonball in the following way: "As the cannonball comes up, the force from the cannon is dissipating and the force of gravity is taking over. So it's slowing down. At the peak the force of the cannon and the force of gravity are fairly equal. That's the changeover point, at which point the force of the cannon is less as gravity becomes the more predominant force. As it makes the arc and begins to come down, gravity is overcoming the force from the cannon." Clement found that students give similar explanations for the motion of a coin tossed upward.

Some subjects who gave this kind of

explanation stated that a projectile accelerates as it falls because of the continuing dissipation of the original upward impetus. Perhaps it bears repeating that according to Newtonian mechanics, there is no upward impetus and no upward force; there is only the upward component of the projectile's velocity, which is steadily reduced by the action of gravity. The peak of the projectile's arc does not represent a point at which upward and downward forces are equal, as it does in the impetus theory.

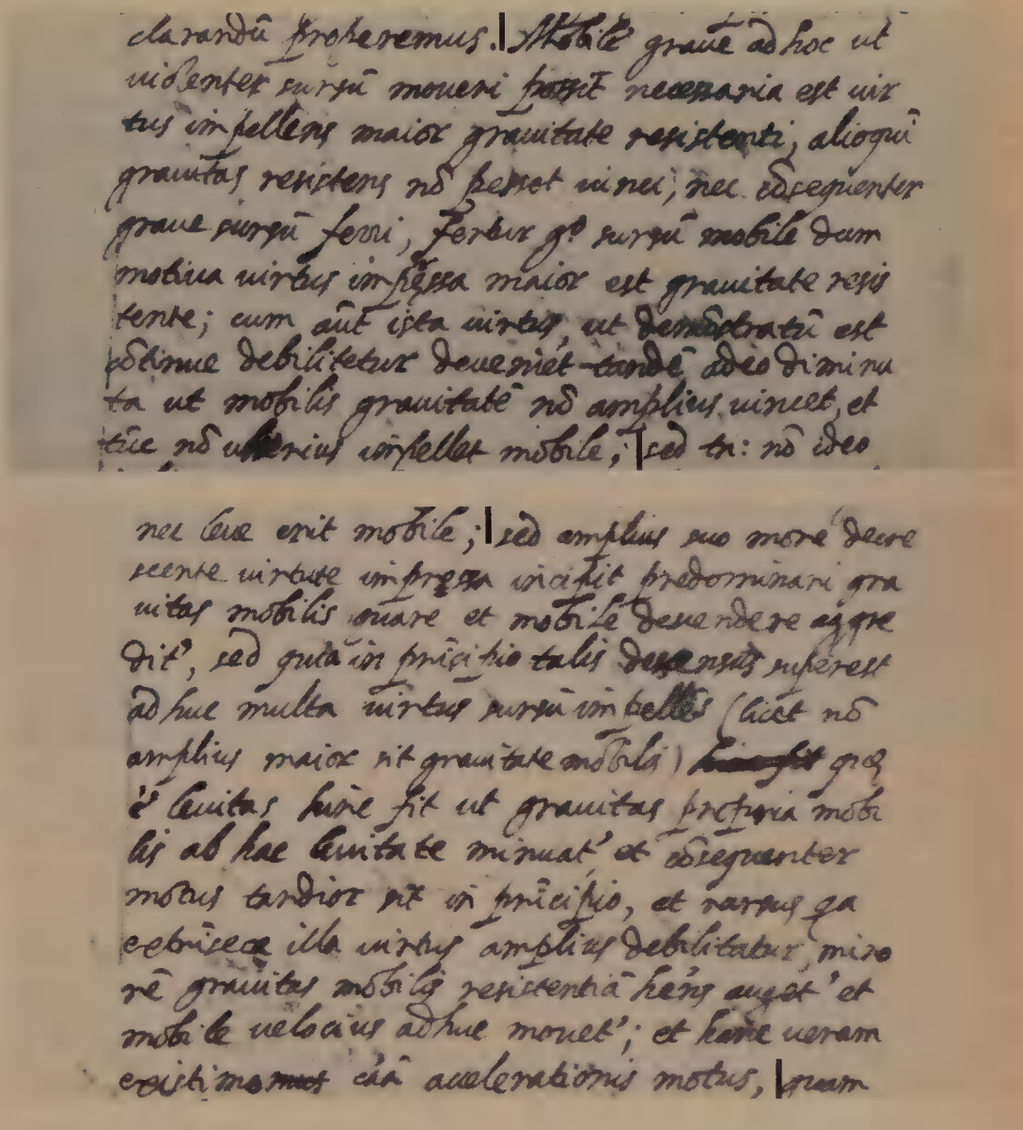
In one of our problems on projectile motion we asked the subjects to suppose a metal ball is pushed along the top of a cliff at high speed and then goes over the edge. Their task was to draw the path of the ball as it falls, ignoring air resistance. The correct answer is that the ball's constant horizontal velocity and its steadily increasing vertical velocity combine to generate a parabolic arc.

More than a third of our subjects who were given the cliff problem unknowingly took the position of Albert of Saxony: the ball, they maintained, travels in a straight horizontal line for some distance beyond the edge of the cliff. It then begins to curve downward and eventually falls straight down. One subject explained: "When [the ball] leaves the cliff, the horizontal force is greater than the downward-motion force. When the horizontal force becomes less, the ball would start falling." Five percent of our subjects echoed Avicenna's theory: the ball, they said, travels straight out and then straight down. One subject who gave this answer defended his reasoning by saying that "gravity isn't going to affect it until it stops moving."

The belief that a carried object falls straight to the ground when it is dropped also derives from the intuitive impetus theory. Subjects who hold this belief explain that an object acquires impetus when it is pushed or thrown but not when it is passively carried. Therefore, according to the impetus theory, when a carried object is dropped, there is no force to cause it to move forward, and so it falls straight down. The intuitive belief is maintained even for situations where the carrier is moving at a high speed. In one experiment we asked students to draw the trajectory of a bomb dropped from a moving airplane, ignoring the effects of air resistance. More than a third of the subjects indicated that the bomb would fall in a straight vertical path.

The striking similarity between the views of the medieval philosophers and those of our subjects suggests the impetus theory is a natural outcome of experience with terrestrial motion. It is easy to see how experience could lead to the basic idea that objects are kept in motion by an internal force that gradually dissipates: in most circumstances an object that is set in motion keeps moving for some time, but it eventually stops. What is difficult to understand is why people develop incorrect beliefs about the trajectories of moving objects, beliefs that apparently conflict with everyday experience. Why, for example, do people come to believe carried objects fall straight down when they are dropped?

A possible answer is that under some conditions the motion of objects is systematically misperceived. Washburn, Linda Felch and I have proposed that objects dropped from a moving carrier are often perceived as falling straight down. The misperception may be the source of the corresponding belief that they do so. Studies in the perception of motion have shown that when an object is viewed against a moving frame of reference, a visual illusion can arise. The motion of the object relative to the moving frame of reference can be misper-

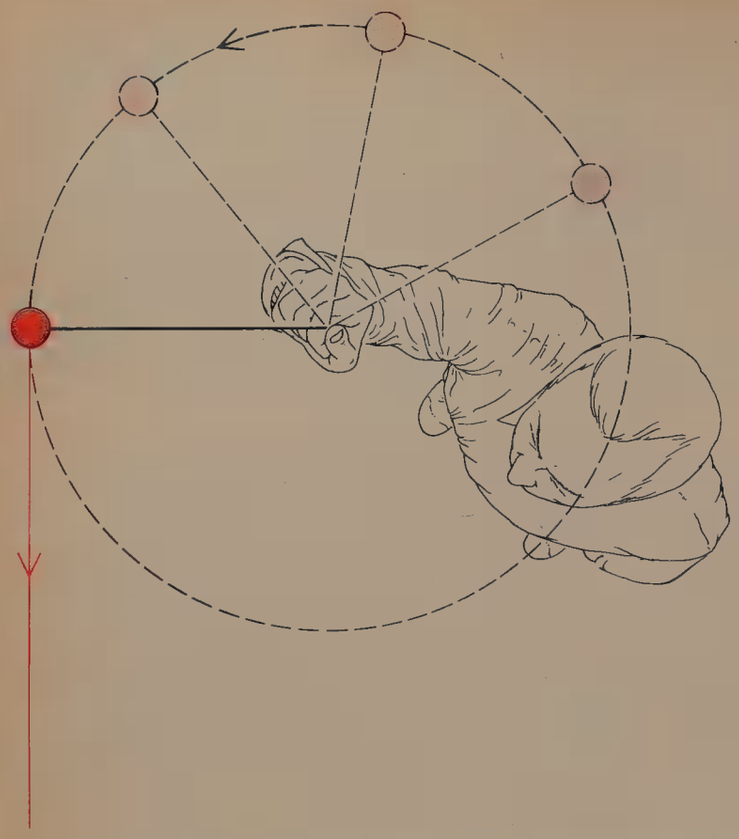


For a heavy body to be able to be moved upward by force, an impelling force greater than the resisting weight is required; otherwise the resisting weight could not be overcome and consequently the body could not move upward. That is, the body moves upward, provided the impressed motive force is greater than the resisting weight. But since that force, as has been shown, is continuously weakened, it will finally become so diminished that it will no longer overcome the weight of the body and will not impel the body beyond that point. . . .

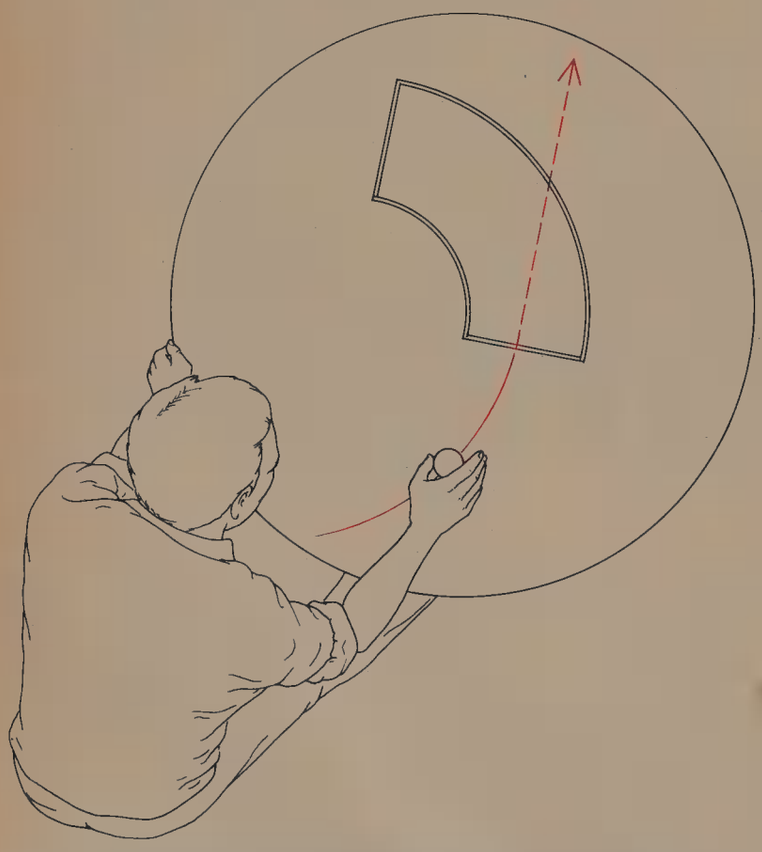
But beyond that, as the impressed force characteristically continues to decrease, the weight of the body begins to be predominant and consequently the body begins to fall. Yet there still remains, at the beginning of the descent, a considerable force that impels the body upward, which constitutes lightness, although the force is no longer greater than the heaviness of the body. For this reason the essential weight of the body is diminished by the lightness and consequently the motion is slower at the beginning. Furthermore, since that external force continues to be weakened, the weight of the body, being offset by diminishing resistance, is increased and the body moves faster and faster.

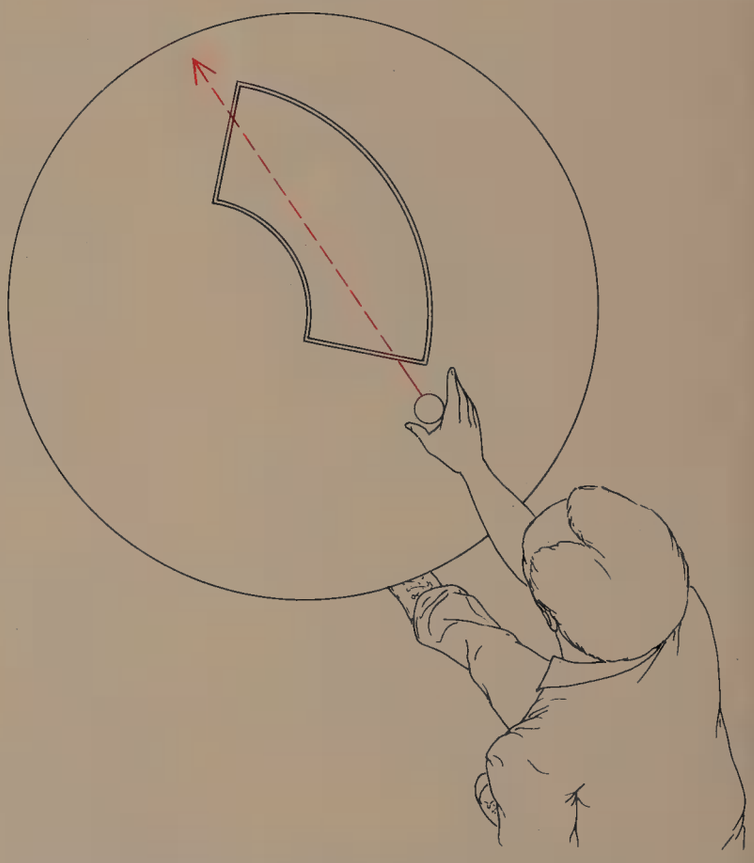
This is what I consider to be the true cause of the acceleration of motion.

MANUSCRIPT OF GALILEO'S *DE MOTU* ("On Motion"), an essay written in about 1590, shows that he endorsed the impetus theory in his early writings. In the excerpt Galileo applies the concept of impetus to explain the motion of ■ projectile launched upward. The author and his colleagues at Johns Hopkins University and John J. Clement of the University of Massachusetts at Amherst have found that experimental subjects who are asked to explain the rise and fall of a projectile frequently give quite similar accounts. The manuscript is in the Biblioteca Nazionale Centrale in Florence; the English translation was done by the late I. E. Drabkin.



CIRCULAR IMPETUS, according to the impetus theory, is imparted to an object when the object is moved in a circle. Hence the impetus theory incorrectly predicts that a ball whirled at the end of a string will continue to follow a curved path if the string breaks. Actually the ball will move in a straight line along a tangent to its original circular path, beginning at the instant the string breaks. Fifty-one percent of the college students who were asked to draw the path of the ball after the string breaks sketched the path correctly (*left*). Thirty percent, however, thought circular motion would persist and the path would be curved (*right*); 19 percent gave other incorrect responses.





SOME PEOPLE INTERACT WITH MOVING OBJECTS as if the objects could be given circular impetus. A 90-degree segment of a ring was painted on a table, and experimental subjects were given a small "puck" with a ball bearing that would enable it to roll smoothly across the table. The subjects were asked to push the puck up to one edge of the ring segment and release it; the task was to make the puck cross to the other side of the segment without touching the curved sides. Twenty-five percent of the subjects tried the strategy at the left: they moved the puck in an arc, apparently in the mistaken belief that the object would continue to travel in a curved path as it moved through the ring segment. The broken line, which represents the path of the puck after it was released, shows that this strategy invariably failed. Sixty-seven percent of the subjects applied the correct strategy that is shown at the right: they aimed the puck so as to take advantage of its straight-line trajectory after it was released. The remaining eight percent of the subjects tried other unsuccessful strategies.

ceived as absolute motion (that is, as motion relative to the stationary environment). For example, if a dot in the interior of a rectangle remains motionless as the rectangle moves to the right, the dot can be perceived as moving to the left [see "The Perception of Motion," by Hans Wallach; SCIENTIFIC AMERICAN, July, 1959].

Similarly, when people observe carried objects that are dropped, the moving carrier usually remains in view as the dropped object falls; hence the carrier can act as a frame of reference. In situations where the effects of air resistance are negligible, the dropped object falls straight down with respect to the carrier. Therefore the illusion could develop in which the object appears to fall in a straight vertical trajectory with respect to the ground.

In order to test this hypothesis we asked subjects to watch computer-generated displays resembling situations in which carried objects are dropped. A rectangular box, representing the carrier, moved across a video screen at a moderate, constant speed from left to right. A round dot, representing the carried object, remained at the top center of the moving box until the box had traveled about a third of the distance across the screen. The dot then traced one of several paths that were designed to simulate the trajectory of an object dropped in the presence or absence of air resistance. The box kept moving at its initial speed while the dot fell. For each display the subject first watched the motion and then tried to draw the path that had been taken by the dot.

Our results indicated that subjects experienced a strong illusion when they viewed the displays. When the dot traced the path of an object falling in a vacuum, our subjects generally drew a path that was much steeper than the one they had just observed. That is, they perceived the path of the dot as being nearly vertical and curved only slightly in the direction of the motion of the box, even though the dot had actually moved much farther to the right. When the dot fell like an object affected slightly by air resistance, it was generally perceived as falling straight down, although again it actually moved considerably farther to the right during its fall. The illusion persisted only when the moving box was shown; when the subjects observed the same motion of the dot in the absence of the moving frame of reference, the illusion did not arise and the drawings were accurate.

In a second experiment our subjects watched a videotape of a walking person dropping a ball; again their task was to draw the path of the ball. We warned them that the videotaped event might have been faked through trick photogra-

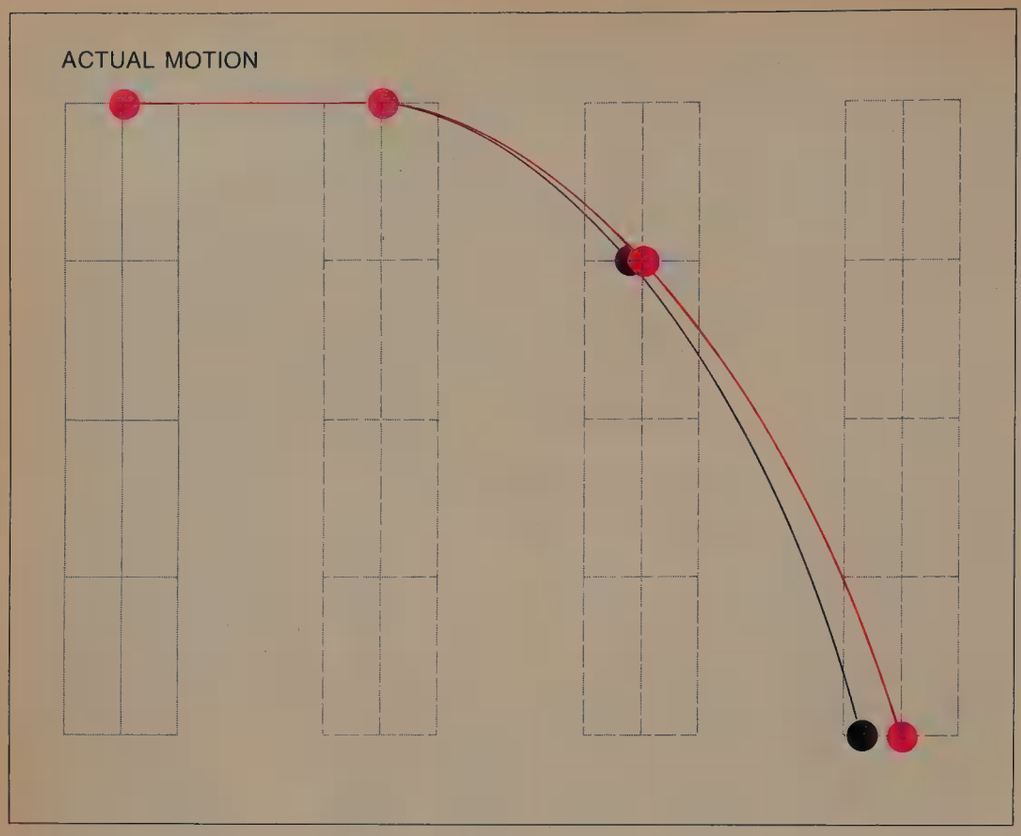
Write us if you'd like a booklet describing what you'll see on the distillery tour . . . rain or no rain.

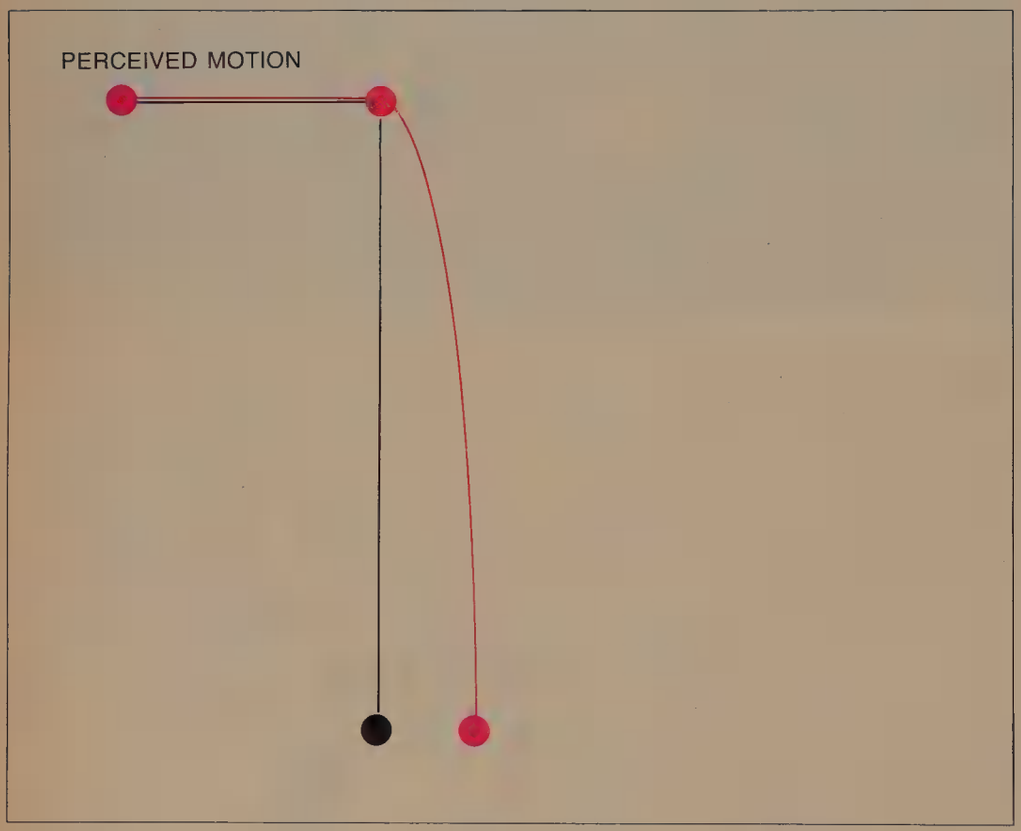
SOME DAYS, visitors to Jack Daniel's are surprised to hear they're in a dry county.

It's "dry" because we aren't allowed to sell (or drink) our whiskey here. But as everyone knows, we make a good deal of it. And we enjoy taking folks from one end of our hollow to the other to show them how it's done. Of course, there's no guaranteeing perfect weather. But if you visit our distillery sometime soon, we're certain you'll have a nice day.

CHARCOAL MELLOWED DROP BY DROP

Tennessee Whiskey • 90 Proof • Distilled and Bottled by Jack Daniel Distillery
Lem Motlow, Prop., Inc., Route 1, Lynchburg (Pop. 361), Tennessee 37352
Placed in the National Register of Historic Places by the United States Government.





ILLUSIONS IN THE PERCEPTION OF MOTION may explain some of the widespread misconceptions people have about moving objects. The author and two of his students, Allyson Washburn and Linda Felch, tested the relevance of this hypothesis for explaining the incorrect belief that an object dropped by a moving carrier falls straight down. A rectangular box, representing the carrier, moved from left to right across a video screen. A dot, representing the carried object, remained fixed at the top center of the box for about a third of the distance across the screen and then began to fall. The falling dot moved like an object dropped in a vacuum (*colored curve in upper illustration*) or like an object affected slightly by the resistance of the air (*black curve in upper illustration*). After watching the motion of the dot and the rectangle experimental subjects tried to draw the path the dot had just traced with respect to the screen. When the dot fell as if it were in a vacuum, it was generally perceived as moving only slightly forward in the course of its fall (*colored curve in lower illustration*). When the dot moved as if it were influenced by the resistance of the air, the average perceived path was a straight vertical line (*black curve in lower illustration*). There was no such illusion when subjects viewed the same motions of the dot in the absence of the moving rectangular background.

phy, although actually it was not. Consequently, we told them, they were to draw exactly what they observed and not what they thought should have happened. Of the 18 subjects who viewed the videotape only four indicated that the ball moved forward as it fell. Ten subjects said the ball fell straight down and four thought it moved backward.

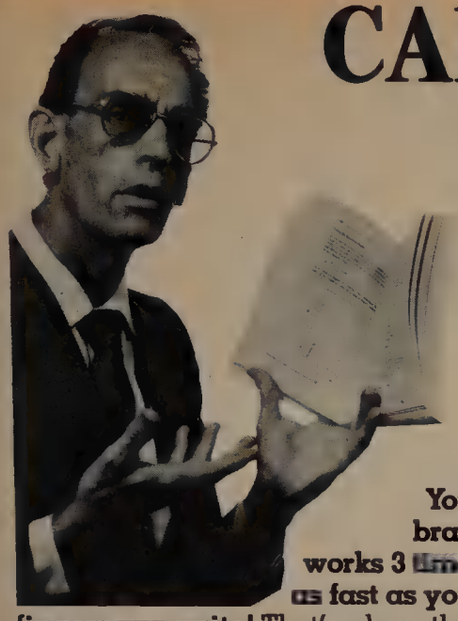
Such results do not prove that misperception is the cause of mistaken beliefs about the motion of dropped objects, but they do suggest strongly that the hypothesis is a reasonable one. My colleagues and I are exploring the possibility that illusions are implicated in the development of other misconceptions about motion.

How can such misconceptions be dispelled? The most obvious answer is through normal instruction in Newtonian mechanics, but studies by several investigators suggest that the intuitive ideas are difficult to modify. Although some students who take physics courses achieve a firm grasp of Newtonian mechanics, many emerge with their intuitive impetus theories largely intact. Kohl, Washburn and I tested the knowledge of motion among high school students before and after they had finished a physics course and found that the course had left some misconceptions unaffected. The instruction was particularly unsuccessful in altering the core assumption that impetus acquired when an object is set in motion serves to maintain the motion: 80 percent of the students retained this belief even after finishing the course. (Ninety-three percent held the belief prior to instruction.)

It appears that students often rely on the intuitive impetus theory as a framework for interpreting new course material. As a result the material can be misinterpreted and distorted to fit the intuitive preconceptions. Several educators are seeking ways of designing physics instruction to avoid such pitfalls. James Minstrell, a high school teacher in Mercer Island, Wash., has reported some success with a method whereby students are led to articulate their intuitive beliefs. The differences between such beliefs and the principles of Newtonian mechanics can then be sorted out. Work by Andrea A. di Sessa of the Massachusetts Institute of Technology suggests that experience with computer games, in which objects act as if they moved in a frictionless Newtonian world, may also be helpful.

Hence the study of intuitive theories and the processes by which they are acquired and modified holds promise for the development of improved educational methods. Such study may also help to illuminate the factors that influence the historical development of formal scientific theories.

CALCULATOR MATH?



Your brain works 3 times as fast as your fingers can write! That's where the Calculator comes in — you can go three times as fast with your machine doing the dirty work.

It's easier and more interesting. And you'll understand it better when it's fun. (If you haven't yet tried out a modern scientific-statistical calculator, you're in for a pleasant surprise — see below!)

WHY MATH? To get a promotion or a new job. Just to hang on to that job you've got. To quicken your mind. To understand this world you live in. To impress people. Just for some fun you'll feel good about. (Read those letters below to see how it has worked for others.)

Do it at home... You can start right away and go as fast as you like! Put it aside while you're busy with something else — in short, *set your own pace!*

Lots of us like to avoid being compared with the other members of a class — we like to set our own standards. Too, we like to save the time and expense of going to a school or seminar. (Best of all, it's a lot cheaper to study at home where the rent is already paid — *read on!*)

Let's be honest. You wouldn't expect to learn to play baseball or chess by watching an expert do it. Math is the same — you'll need to do some work. BUT — with my calculator method, you'll see it's a whole new ball game!

THIS METHOD IS FRIENDLY! You can use it on your own. You follow interesting examples step-by-step with calculator keystrokes — instead of algebra.

Let me introduce myself — I'm Dr. George McCarty and I teach at the University of California. Tens of thousands of people like yourself have already used my guidebook to teach themselves calculus. Recently I have teamed up with another professor. We put those same tested and proven techniques to work for you in a new *Statistics Guidebook.*

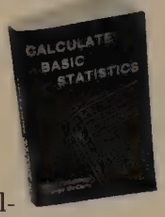
Even if you've had trouble with math, this method will work for you. Choose the one you want to learn, calculus or statistics (or both) and start *today!*

As pennywise Ben Franklin said, "An investment in knowledge pays the best dividends."

CHOOSE STATISTICS

It's used everywhere — to set insurance rates and to assess a drug's effectiveness, to detect a crooked roulette wheel and to measure lightbulb lifetimes, to audit the books and to find your position at sea.

People have always known that their experience was not a perfect guide to the future. Statistics shows us that uncertainty is not only inevitable but it itself is measurable.

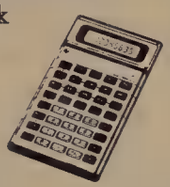


You'll find out about this in my guidebook. You'll read informal explanations with detailed calculations for topics like "confidence intervals" and "contingency tables."

YOU'LL BE DOING Inference and Hypothesis Testing, the Chi-Square Distribution, ANOVA, Regression and Correlation, Non-Parametric Methods — they're all here.

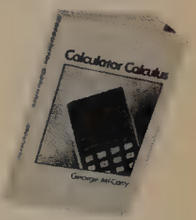
SUPPOSE YOU HAVE ALREADY STUDIED STATISTICS. You'll use Quick-Reference Guides to go straight to a topic you want to refresh yourself on — then get *right-now* answers in your own field, business, biology, science, psychology, education.

The Guidebook is guaranteed to work with any calculator or computer. If you would like to start out with a fancy new machine, order my Statistics Kit which includes the guidebook and a TI-35 calculator which finds means and standard deviations at the touch of a button. It has an LCD display, 1000 hour batteries — the works.



CHOOSE CALCULUS

The math of science and engineering — satellite orbits and compound interest, bacterial growth and radioactive decay, marginal profits and reaction rates.



You'll find out about these concepts when you use the Calculus Kit. This Kit contains my guidebook and a TI-35 calculator which finds log and trig values at the touch of a button. (It's modern, too, with LCD display and 1000 hour batteries.)

In the guidebook you'll learn both differential calculus and integral calculus — *even differential equations!*

NEED TO EVALUATE functions, areas, volumes — solve equations — use curves, trig, polar coordinates — find limits for sequences and series? It's all here!

If you're in the biological, social or physical sciences, you'll be doing Bessel functions, carbon dating, Gompertz' growth curves, half-life, future value, marginal costs, motion, cooling, probability, pressure — and plenty more.

Important numerical algorithms are here, too: rational and Padé approximation, bracketing, continued fractions, Euler's method, Heun's method, iteration functions, Newton's method, predictor-corrector, successive substitutions, Simpson's method and synthetic division.

If you already have a scientific calculator, you can get just the *Calculus Guidebook.*

George McCarty

P.S. — YOU CAN DO IT! I'm certain I can show you this easy way to success with math, so *I guarantee it to you unconditionally* — if you are not satisfied just send it back for a prompt refund!

PHONE MC or VISA ORDERS TO
(714) 831-2637
Ask for our Catalog of books and calculators.

LOOK AT WHAT THEY SAY...

- "*...excellent diagrams and clear explanations.*" —POPULAR SCIENCE
- Tektronix Engineer Bill Templeton says, "*The best, most clearly written book I have seen for improving your math skills.*"
- "*...straight-to-the-point explanations...*" —HP Key Notes
- Professor H.I. Freedman of the U. of Alberta, states: "*There can be no question as to the usefulness of this book...lots of exercises...very clearly written...easy reading.*"
- S.C. McCluney, Jr. of Philadelphia writes: "*With your book and a calculator the whole idea becomes clear in a moment, and is a MOST REFRESHING EXPERIENCE.*"

TAX DEDUCTIBLE FOR PROFESSIONALS

YES, I want to learn math **fast** and **easy** with your *new method.* If I am not more than satisfied I can return it for a quick refund. RUSH me the guidebook or complete kit I select below (kit contains the guidebook and a modern TI-35 calculator).

☐ Send **FREE** Catalog of books & calculators

STATISTICS GUIDEBOOK		KIT
How many?	☐ $14.95	☐ $39.95
Handling/Shipping	1.00	2.00
Tax (within Calif.)	.90	2.40
TOTAL		

CALCULUS GUIDEBOOK		KIT
How many?	☐ $19.95	☐ $44.95
Handling/Shipping	1.00	2.00
Tax (within Calif.)	1.20	2.70
TOTAL		

For AIR SHIPMENT add $3.00 per item

Phone (714)831-2637 (MC or VISA only)

Enclosed is ☐ Check, ☐ M.O.; or charge my MC/VISA... exp._____

Card #_____

Name _____

Address _____

City_____

State, Zip _____

OUTSIDE USA, surface shipping is $2.00 per book or $5.00 per kit. Air shipment is $5.00 per book or $10.00 per kit. U.S. FUNDS PLEASE.

EduCALC Publications, Dept. A3
27953 Cabot Road
South Laguna, CA 92677

©1982

THE AMATEUR SCIENTIST

How lenses can be made out of ice, and what happens when coffee is brewing in the ibrik

by Jearl Walker

A lens made out of ice seems at first thought to be improbable. It is quite practical, however, and offers a number of avenues for experiment. The notion figured prominently in one of Jules Verne's science-fiction tales, *The Desert of Ice.* Now it has been brought to reality by Matthew G. Wheeler of McBride, B.C.

In Verne's story a group of explorers survive a catastrophe in the Arctic by means of an ice lens. The ship's captain had set out to reach the North Pole, but a mutiny had left him and a few supporters with a broken and sinking ship and few supplies. Their only chance to survive was to reach another ship that had been abandoned when it became locked in the ice. The group gathered wood from their broken ship, packed up their meager food supplies and the flint and steel for making fire and set out.

Because their food was running out, they stalked and killed a polar bear, but as they were doing so their fire went out and the flint and steel were lost. Although they had enough food for the remainder of the trip, they were threatened with freezing to death if they could not restart the fire.

They were saved by optics. Noting the bright sunshine, a physician in the group recalled how a fire can be started with a lens. Lacking a glass lens, he decided to make a lens of ice. He chose a clear section of ice and hacked out a piece about a foot in diameter. Shaping it roughly with a hatchet, he whittled and smoothed it with his knife and then polished the surface with his bare hands "until he had obtained as transparent a lens as if it had been made of magnificent crystal." The ice lens was positioned above some tinder, focusing the rays of the sun. Within a few seconds the tinder had caught fire. "The stove was soon roaring, and it was not many minutes before the savory odor of broiled bear-steak" had roused the men.

What the physician had made was a convex lens. Since it was a large one, it intercepted a goodly amount of sunlight. The concentration of the light on the tinder placed at the focal point of the lens quickly delivered enough energy to kindle a fire.

I told this story on one of my regular contributions to Quirks and Quarks, a radio show about science on the Canadian Broadcasting Corporation. I also described how one can fashion a convex lens of ice in a simpler way with curved watch glasses. (A watch glass is the kind of glass that covers the face of a watch. Glasses of similar shape but larger size can be found among the standard supplies of a chemistry laboratory.) Water is poured into two watch glasses and allowed to freeze. Then the flat sides of the two pieces of ice are placed together to form the lens.

Wheeler was a listener who decided to improve on my idea. He made a lens from the ice in a cattle trough. He started with a section that was about an inch thick and free of bubbles. To shape the piece he put it in the lid of a milk pasteurizer, which is curved like a lid for a saucepan. Then he floated the lid (with the piece of ice) in a sink of warm water. The heat melted the bottom of the ice until it fitted the inside of the lid. He occasionally rotated the ice to increase the symmetry of its bottom surface. Next he turned the piece over to shape the other side.

Wheeler now had a convex lens of ice. To test its focusing power (and to mimic the physician in the Verne story) he held the lens by its edge and focused the light of the sun onto a newspaper: "The paper quickly singed, smoked and burst into flame." Wheeler tells me the lens worked best when the air temperature was above freezing. At lower temperatures drops of water condensed on the lens, distorting its surface and interfering with its focusing of light.

Wheeler made a smaller lens in the concave base of a pressurized container. When he placed the lens in front of his 35-millimeter camera, it functioned as a telephoto lens with a focal length of about 80 millimeters. Wheeler tested it outdoors on a cold day with snow on the ground. The lens was too fast to photograph snow scenes, and so he stopped it down to $f9$ with a cardboard iris. The iris was mounted between a collar of black plastic tubing and a conventional rubber lens shade. The purpose of the collar was to eliminate any light that did not pass through the lens.

Initially Wheeler held the lens in place with his gloved hand. Later he found that a cardboard mount was more convenient. The lens was clamped to it with three cardboard flanges stapled to the card. On the front of the card he put a cardboard shade. Once the lens was mounted in its holder the card and the lens were held against the rubber lens shade and adjusted for position. Wheeler told me an ice lens will work well for 15 minutes and will serve even when the air temperature is 16 degrees Celsius (about 60 degrees Fahrenheit).

Wheeler has now made many lenses with a variety of molds. In some cases he begins with liquid water that he freezes. He says the clearest ice results when the freezing begins at the top. To make this kind of ice he insulates the sides and the base of his mold.

If you make an ice lens with a pair of watch glasses, experiment with placing the two pieces of ice together to complete the lens. The pieces should have good contact. Try rubbing them together or stroking them with a finger so that the contact surfaces are wet and smooth. When they freeze together, the contact should be firm.

By choosing appropriate molds you can fashion both convex and concave lenses or combinations of the two. The index of refraction for ice is less than that for glass, so that an ice lens does not bend the rays of the sun as much as a glass lens of the same shape. Therefore an ice lens (either convex or concave) has a longer focal length than a similar lens made of glass.

Middle Eastern coffee is a strong drink brewed from sugar and finely ground coffee. The sugar is needed to offset the bitter taste of the coffee, but I think it also is a factor in the brewing. A mixture of ground coffee, sugar and water is heated in an ibrik, a container that tapers slightly toward the top and has a long handle. The coffee is rapidly brought to a boil and then poured into a demitasse, which is half the size of a standard coffee cup. What is poured is a mixture of liquid, foam and grains of coffee. The foam is highly regarded in Middle Eastern countries. In fact, some people say it is an insult to serve a guest a demitasse of coffee without foam.

One must let the coffee remain in the demitasse for several minutes before it is tasted. The grains of coffee settle slowly to the bottom. If the brew is sampled too soon, the liquid is unpleasantly gritty. After the grains have settled the liquid is sipped slowly until nothing

Matthew G. Wheeler employing one of his ice lenses to ignite paper

A scene in British Columbia photographed by Wheeler with an ice lens mounted on a camera

remains but the bottom layer of coffee grains, which is about the consistency of a sludge.

I have spent hours meditating on the brewing and sedimentation of Middle Eastern coffee as I prepare demitasse after demitasse of the drink. It seems to be about the only coffee drink in which grains of the coffee remain in the liquid after the brewing process. It is also noteworthy in several other respects. Why must the coffee be finely ground? (When you buy coffee for the drink, you should ask for an extra-fine grind.) Why is sugar heated with the coffee rather than added later as it usually is with other coffee drinks? Why does the liquid suddenly foam up the sides of the ibrik as the coffee is being brewed? How long must one let the coffee stand in a demitasse to ensure an ungritty sip?

I buy finely ground coffee roasted by the Viennese or the French method. Both coffees are darker than the others commonly sold in the U.S., meaning that they have been roasted longer and have lost more of the volatile substances in them. Nevertheless, I doubt that the type of coffee alters anything but the taste of the drink.

I prepare my coffee in a tin ibrik large enough to fill two demitasses, each of which holds about two fluid ounces (roughly 60 milliliters). Usually, however, I make only a single demitasse at a time, otherwise the foam forms so fast that it cascades onto the stove. Hence I add a bit less water to the ibrik than is needed to fill one demitasse. I add two teaspoons of coffee grounds and from one teaspoon to four teaspoons of sugar to the water. (One teaspoon is equivalent to five milliliters.)

Some people mix the ingredients before brewing the drink, but I usually do not. I put the ibrik on the heating coil of my electric stove with the control set on high. Soon the mixture begins to rumble and click. After several minutes the noise suddenly stops and the foaming starts. The foam rapidly builds upward to the top of the ibrik. The transition is so fast that I sometimes fail to remove the ibrik from the heating coil before the foam overflows. If I pour the coffee slowly into the demitasse, some of the foam remains in the cup, but most of it collapses.

In order to look inside the mixture while it was brewing I replaced my ibrik with a 180-milliliter beaker made of Pyrex. Again I added the proper amount of water and coffee for one demitasse. The coffee grounds floated and very slowly became wet. Apparently they are held together with enough electric force to keep them in a clump. Water seeps between the grains only slowly. Although the grains are denser than water and so should sink, the clumps contain enough trapped air to make them float.

I put two teaspoons of sugar onto the coffee. Most of the sugar sank, sometimes tipping over a clump of coffee as it did so. When I placed the beaker on the heating coil, the sugar and coffee layers were well separated by a layer of relatively clear water.

Even if I stir the mixture before heating it, the same three layers appear. In cold water the sugar dissolves slowly and so settles to the bottom after being stirred because it is denser than water. Even with stirring, the coffee grains still carry enough adsorbed air to make most of them float to the surface. The only difference when the mixture is stirred is that some of the grains end up in the water layer.

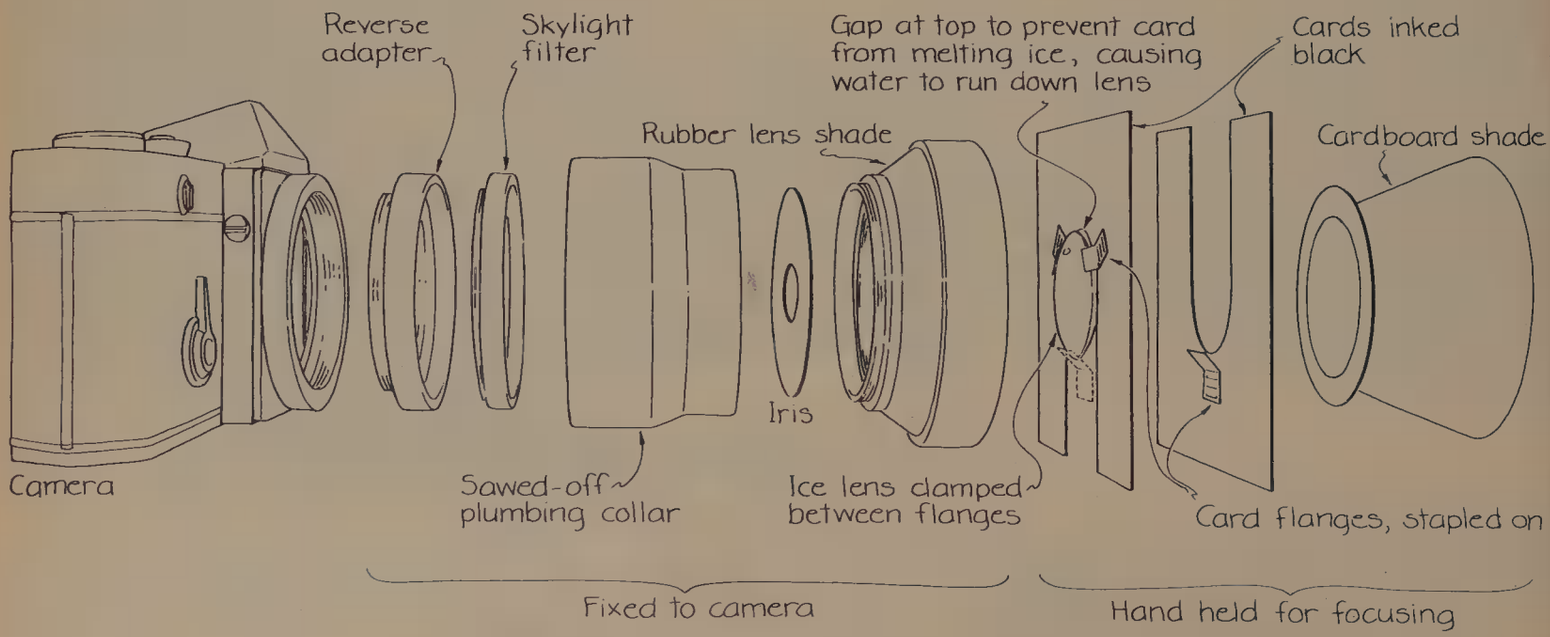
Soon after the heat was turned on bubbles of air appeared at several sites along the bottom of the sugar layer and began oscillating rapidly. The air in the bubbles had been adsorbed onto the sugar grains or had been dissolved in the water between the grains. Once bubble nucleation starts at a given place more air is added to the site from the surrounding water and sugar. The bubbles oscillate because when they expand upward into cooler liquid (concentrated sugar water in this case), they suddenly collapse. Soon afterward, however, they are reheated and reexpand. The rapid and chaotic oscillation of the air bubbles generates the clicks and rumbles characteristic of the early stage in heating a fluid.

Bubbles of air and water vapor eventually break free and stream upward through the sugar layer. When they penetrate the top of the layer and enter the clear water, their motion is arrested. The water is cooler than the sugar layer and makes the bubbles collapse.

By shining a flashlight into the beaker I can spot threads of sugar water projected upward from the collapse of each bubble. The threads are visible because their index of refraction differs from that of the water. The threads twist and turn before they fall back to the sugar layer. Sometimes small droplets of sugar (or concentrated sugar water) are left in the wake of the threads.

Soon the entire sugar layer becomes turbulent. Whereas until then bubbles conveyed heat upward to the layer of water, convection streams now seem to dominate. The variation in the index of refraction makes some of these streams visible. With further heating the turbulent sugar layer moves upward into the water layer. Some of the turbulence or one of the bubbles from the top of the sugar layer occasionally reaches the bottom of the coffee, but otherwise the coffee seems to be unaffected by what is taking place below it.

As the turbulent layer of sugar nears the coffee the grains begin to glisten with



Wheeler's arrangement for mounting an ice lens on a camera

tiny air bubbles, none larger than a millimeter and most much smaller. Just as the sugar layer reaches the coffee the grains are stirred around and yield a great many of the tiny bubbles. They form the foam, which builds up rapidly. Within a few seconds the foam is three times as high as the initial mixture.

As soon as the grains are stirred around the rumbles and clicks stop. The noise originated in the formation of large bubbles along the bottom of the container. Once the grains reach the region they provide plenty of nucleating sites for smaller bubbles. The large, noisy bubbles disappear.

I employ a long pair of tongs to grasp the hot beaker. If I hold the beaker above the heating coil, I can control the rate of production of the foam: the rate slows as I move the beaker away from the coil. The bubbles are small at first, but as they are pushed up the beaker by newly made bubbles they merge and give rise to larger bubbles. The larger bubbles are too fragile to survive when the coffee is poured. The foam some people like to have on their Middle Eastern coffee consists of the tiny bubbles that were adhering to grains of the coffee when the foaming started. When the coffee is poured into a demitasse, some of the tiny bubbles break free of the grains and form a raft that may hold its shape for quite a while.

I do not always get such a raft on my coffee. When I do, it can survive for hours if the demitasse is left undisturbed. One reason for the stability of the bubbles is that the sugar in the coffee increases the viscosity of the fluid. When a bubble raft has formed, the fluid in the walls between the pockets of gas drains because of gravity. The walls get thinner until the bubbles burst or merge. The more viscous the fluid is, the slower it drains. The fluid in the walls of the bubbles on my coffee drains very slowly.

To change the conditions I made the coffee without the sugar. Again the coffee grains floated in clumps on top of the water. When the beaker was placed on the coil, the water began to heat rapidly and with much noise. Air bubbles appeared and oscillated. Then bubbles of water vapor formed and escaped upward to beat against the bottom of the coffee layer.

Soon streams of these bubbles broke the clumps of coffee. The grains quickly dispersed throughout the volume of the fluid. Although the clumps lost some of their air and thus their buoyancy, the grains still bore adsorbed air. They were stirred down from the top surface only because of the strong convection currents in the beaker. Small bubbles formed and then the foaming began. All these things happened faster than they did when I had sugar in the mixture.

I tried several other preparations. In one I heated only sugar (two teaspoons) and water. As before, the sugar sank to the bottom and was heated first. Viscous bubbles and turbulent convection appeared again. When the mixture was at full rolling boil, bubbles appeared on the top, but with this preparation there was no rapid foaming. The bubbles were relatively large (several millimeters across) and did not last long.

The equipment and materials needed to make Middle Eastern coffee

After a few minutes the fluid was transparent, indicating that the sugar had dissolved fully and was uniformly mixed. I allowed the mixture to cool. Even at room temperature all the sugar remained dissolved. Clearly the amount of sugar normally added to the drink does not exceed the saturation limit of sugar in hot water.

Next I boiled the usual amount of water in the beaker but omitted the sugar. After the rolling boil began I added a single teaspoon of coffee. Almost immediately the grains were swirled throughout the volume of water and tiny bubbles developed on them. The foaming was so fast that some of the mixture spilled onto the stove before I could remove the beaker from the heat.

In another preparation I replaced the sugar with an equal amount of honey. I poured the honey into the beaker first, then the water and finally the coffee. When the honey began to heat up, it formed tiny geysers that rose into the layer of water. They were visible only when I shined a flashlight into the beaker, making the difference in the index of refraction between the geysers and the water apparent. If a geyser pinches off from the main layer of honey, it falls back to that layer because honey is denser than water. Often the stream loops or releases drops of honey. Some of the drops remain stationary, apparently because they have been diluted enough to acquire about the same density as that of water.

At this point in my investigation I could answer some of my questions about the brewing of Middle Eastern coffee. The principle of the brewing is to cook the grains fast. If they are overcooked, the drink tastes too bitter or acidic. It is then like regular coffee that has been left simmering in the pot for a long time. The grains for Middle Eastern coffee are finely ground so that they can be rapidly heated and cooked.

The foam is almost a side effect. The grains are so small that they provide plenty of surface area for adsorbing air. The tiny pockets of air are sites for the nucleation of bubbles. When the coffee and the surrounding water get warm, air dissolved in the water comes out of solution at the nucleation sites, forming bubbles just barely large enough to be seen.

When the hot liquid layer reaches the grains, water vapor expands the bubbles. They pinch off and break free from the nucleating sites and then form the foam of large bubbles that climbs the ibrik. Many of the nucleating sites on the grains remain active because they retain part of the air or vapor after a bubble has left. Moreover, they are now stirred around in the water near the hot bottom of the container. Hence they continue to generate bubbles as long as the surrounding fluid is hot enough to provide more water vapor.

Sugar is added before brewing for several reasons. Heating ensures that the sugar is fully dissolved by the time the coffee is drunk. If the sugar were added after the coffee had been poured into a demitasse, you would have to stir the sugar into the drink to dissolve it. This action would cause the coffee grains to be suspended for much longer. The coffee would be cold before they had set-

Amateur Telescope Making

Edited by Albert G. Ingalls
Foreword by Harlow Shapley

This three-book set is the authoritative reference library of the enthralling hobby of amateur telescope making. Through these books thousands have discovered a fascinating mechanical art combined with a great science.

BOOK ONE begins at the beginning, teaches the basics of glass grinding and how to complete the first telescope. (510 pages, 300 illustrations.)

BOOK TWO leads on into advanced methods of amateur optical work and describes new projects for the telescope maker. (650 pages, 361 illustrations.)

BOOK THREE opens up further fields of enterprise: binoculars, camera lenses, spectrographs, Schmidt optics, ray tracing (made easy). (646 pages, 320 illustrations.)

SCIENTIFIC AMERICAN ATM Dept.,
415 Madison Avenue, New York, N. Y. 10017

Please send me postpaid the following AMATEUR TELESCOPE MAKING books.
My remittance of $_____
is enclosed.

☐ BOOK ONE	$10.00
☐ BOOK TWO	$12.50
☐ BOOK THREE	$12.50

For U.S. shipments add $1.00 each; elsewhere add $2.00 each.

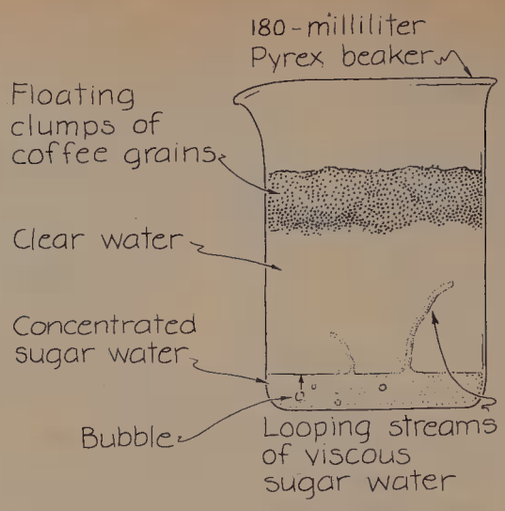
Name_____

Address_____

City_____

State_____ Zip_____

Residents of New York City please add city sales tax. Other NYS residents please add state sales tax plus local tax.



Making the coffee in a beaker

tled out enough to provide a drink that was not gritty.

The sugar also facilitates brewing. The viscous layer of sugar at the bottom of the ibrik holds heat better than water. If the layer consisted of water only, bubbles of water vapor would transfer heat to the coffee layer early in the brewing procedure. With the sugar in place the bubbles and their heat transfer are retarded. Only later does the hot and turbulent layer of sugar water reach the coffee and begin to heat it. The period of heating and cooking of the coffee is sudden and short. It begins when the hot liquid reaches the coffee; it ends in less than a minute when the foam threatens to overflow the container.

Finally, the sugar is beneficial to the brewing process because it helps to stabilize the foam of tiny bubbles. Without the sugar the foam is less likely to survive being poured into the demitasse.

Although the brewing process is fast, it must not be given too little time. When I make enough coffee for two demitasses simultaneously in my ibrik, I often wind up with gritty coffee. No matter how long I wait large grains fail to settle. The reason is that they still carry enough adsorbed air to remain buoyant. They may even remain clumped together. Since the ibrik has too much fluid, I must take it off the stove prematurely to keep foam from overflowing. As a result the time from when the hot, turbulent sugar layer reaches the coffee to when the coffee is poured is too short for all the adsorbed air to be eliminated. From experience I have learned to prepare only one demitasse of coffee at a time. If it is necessary to make a larger amount in one preparation, I must hold the ibrik off the heat source to control the foam and to cook the coffee long enough.

My next set of questions has to do with what is going on in the demitasse as the grains of coffee sink to the bottom. Each grain sinks at a speed that is approximately constant. Since the sedimentation rate of suspended particles is an important datum in both science and technology, much work has been done to develop a theory of how fast particles settle in water.

If the particles are smaller than about .06 millimeter in diameter, they fall according to Stokes's law, named after the British physicist and mathematician G. G. Stokes. The speed of a small particle depends on two factors: the effective weight of the particle and the drag applied to it by the fluid. The effective weight is the difference between the particle's true weight and the buoyancy imparted to it by the surrounding fluid. This difference is usually described in terms of the density of the particle and the fluid. Since a coffee grain is denser than water (and denser than a mixture of sugar and water), the buoyancy imparted to it is too weak to keep it from falling. Its effective weight is a downward force smaller than its true weight.

Countering this force is an upward force generated by drag. According to Stokes's law, the drag depends on the diameter of the particle, the speed of the particle and the viscosity of the fluid. When the particle begins to descend, its speed is small and so is the drag. The particle steadily accelerates and soon is moving fast enough for the upward drag force to match the downward force of the particle's effective weight. Thereafter the particle no longer accelerates, but it continues to fall at a constant speed. This terminal speed depends on the size of the particle. Smaller particles have lower terminal speeds and therefore take longer to settle.

The value of the terminal speed also depends on the viscosity of the fluid. Since the sugar increases the viscosity of the coffee, all the particles have lower terminal speeds than they would have without the sugar.

Unfortunately for drinkers of Middle Eastern coffee, Stokes's law applies only to spherical particles. The coffee grains small enough to be affected by the law are irregular in shape and therefore encounter a drag force that is difficult to calculate. Nevertheless, the overall dependence on particle size generally holds: smaller grains fall slower.

Many of the grains in Middle Eastern coffee are too large to fall according to Stokes's law, which predicts a drag force that depends on the first power of the particle's speed through the fluid. The larger particles get a drag force that depends on the square of the speed. In addition the equation for the force includes an experimentally determined drag coefficient, which takes into account the effect of the particle's irregular shape.

Another factor is that if there are many particles falling through a certain volume of fluid, they change the situation in two ways. They increase the ef-

fective density of the fluid, thereby decreasing the effective weight of a particle and its settling rate. The particles also interact as they pass one another on the way down. In spite of this added complication the same rule usually holds: smaller particles fall slower.

To demonstrate the rule I poured a small amount of freshly brewed Middle Eastern coffee into a tall, transparent jar containing water. I had left the water undisturbed for an hour so that it was still. I added freshly brewed coffee because I wanted grains that had lost most of their adsorbed air. Unbrewed grains would have tended to float.

With the aid of a flashlight I monitored the descent of the particles. Large ones sank quickly to the bottom at a speed of about one centimeter per second. Smaller grains were much slower. Some of the smallest ones appeared to take 10 minutes or more to reach the bottom.

The settling rate of particles in a demitasse is somewhat different from what I saw in my demonstration. The number of particles per unit of volume is large enough to make it likely that the particles interact as they fall. More important, the fluid is hot enough to create circulation cells. Hot fluid from near the bottom rises to the top, cools by evaporation and then descends. The circulation forms what are called Bénard cells, after the French scientist Henri Bénard. In some fluids and under certain heating conditions the cells create a stable geometric pattern on the top surface.

In my coffee the pattern is chaotic and constantly changing. It is best seen when the coffee lacks foam, since foam not only covers the surface but also alters the evaporation rate. I rig a light that reflects off the surface, revealing that over the region of ascending hot fluid tiny drops are suspended just above the surface. They condense from the water vapor when it cools in the air just above the surface. No drops are visible over the regions of descending fluid. Since the drops reflect light, the regions of ascending fluid look slightly white while the regions of descending fluid are the dark color of the coffee. Bénard cells in a cup of coffee with no grains are merely a novelty. In Middle Eastern coffee the circulation in Bénard cells delays the settling of the grains.

The result of all this physics is that one would do well to strike a balance when drinking Middle Eastern coffee, waiting long enough for the larger particles of coffee to settle to the bottom of the demitasse and for the Bénard circulation to decrease but not so long that the drink gets unpleasantly cool. The foam on the coffee makes the compromise easier by insulating the coffee with the vapor trapped in the bubbles, thereby decreasing the cooling rate.

An example of the interesting things that might turn up unexpectedly while you are casually sweeping the deep sky is this supernova in NGC 6946, in a portion of a photograph taken by Hubert Entrop on November 4, 1980, with his Questar. This was 7 days after the official discovery of the nova, as described in Sky & Telescope *in the January, 1981, issue.*

Great news from Questar. . . The Wide-Sky Telescopes

The comet is coming. And Questar is ready for it with two brand-new instruments for all dedicated comet watchers and other deep-sky enthusiasts.

Our emphasis in developing these new designs was on low magnification and optimum field of view—magnification low enough and field of view wide enough to sweep the sky for richfield observing without need of finder or equatorial mount.

We also wanted to dispel the popular misconception that short focal lengths (f/3 to f/5) are essential for effective deep-sky observing and optimum image brightness. Image brightness is a function of aperture and magnification, and has nothing to do with focal length. We continue to use long focal ratios, even in our new wide-sky models. This rules out the comatic aberrations of short-focal-length reflectors and the spurious chromatic effects of short-focal-length refractors, as well as the below-par imagery of short-focal-length eyepieces. Imagine, then, the pleasure of having that familiar sharp Questar® resolution, with no aberrations, available in a telescope with a field of view of 3°.

Specifications of Questar Wide-Sky Telescopes

Questar Wide-Sky 3½ (shown left): aperture 89mm; focal length 700mm; focal ratio f/7.8; magnification 22 × with 32mm wide-angle eyepiece; field of view 3°; faintest visible star 12th magnitude (visual), 12.8 with special coatings.

Questar Wide-Sky 7: aperture 178mm; focal length 2400mm; focal ratio f/13.5; magnification 44× with 55mm wide-angle eyepiece, 75× with 32mm wide-angle eyepiece, field of view 1.15° with 55mm eyepiece, 1° with 32mm wide-angle eyepiece; faintest visible star 13th magnitude (visual) with standard coatings, 14th magnitude (visual) with special coatings.

Both instruments can be equipped with various accessories to provide for tracking and photography.

Special uses for the Wide-Sky Questars are comet seeking and general sky scanning, monitoring telescopic meteors, observing the occultation of stars and planets by the Moon, observing lunar and solar eclipses, observing the deep-sky phenomena — nebulae, star clusters and galaxies — and variable star observing.

Photographic accessories are available for use in these applications, in which the Questar becomes an astrographic camera. All Questars, of course, can be used with a video monitor.

Let us send you our price list.

© 1983 Questar Corporation

QUESTAR

Box C, Dept. 205, New Hope, PA 18938
(215) 862-5277

QUESTAR, THE WORLD'S FINEST, MOST VERSATILE TELESCOPE, IS DESCRIBED IN OUR BOOKLET IN COLOR, WITH PHOTOGRAPHS BY QUESTAR OWNERS. SEND $2 TO COVER MAILING COSTS ON THIS CONTINENT, BY AIR TO SO. AMERICA, $3.50; EUROPE AND NO. AFRICA, $4; ELSEWHERE, $4.50.

BIBLIOGRAPHY

Readers interested in further explanation of the subjects covered by the articles in this issue may find the following lists of publications helpful.

METAMAGICAL THEMAS

RECURSIVE FUNCTIONS. Rózsa Péter. Academic Press, 1967.

ANATOMY OF LISP. John Allen. McGraw-Hill Book Company, 1978.

ARTIFICIAL INTELLIGENCE PROGRAMMING. Eugene Charniak, C. K. Riesbeck and Drew V. McDermott. Lawrence Erlbaum Associates, Inc., 1980.

HISTORY OF LISP. John McCarthy in *History of Programming Languages,* edited by Richard L. Wexelblat. Academic Press, 1980.

A NUCLEAR-WEAPON-FREE ZONE IN EUROPE

NUCLEAR-WEAPON-FREE ZONES. Report from Vantage Conference 5, Muscatine, Iowa, October 7–9, 1975. The Stanley Foundation, 1975.

COMPREHENSIVE STUDY OF THE QUESTION OF NUCLEAR-WEAPON-FREE ZONES IN ALL ITS ASPECTS. Special Report of the UN Conference of the Committee on Disarmament. United Nations, 1976.

AMERICAN FOREIGN POLICY AND NUCLEAR-WEAPON-FREE ZONES: ISSUES AND OPTIONS. Joseph M. Grieco in *Potomac Review,* Vol. 7, pages 12–23; Winter, 1976.

ARMS CONTROL: THE REGIONAL APPROACH. Ronald G. Purver. Center for International Relations, Queens University, Kingston, 1981.

STUDY ON ALL THE ASPECTS OF REGIONAL DISARMAMENT. United Nations Centre for Disarmament. Report to the Secretary General, United Nations, 1981.

COMMON SECURITY: A BLUEPRINT FOR SURVIVAL. Independent Commission on Disarmament and Security Issues. Simon and Schuster, 1982.

SILICON MICROMECHANICAL DEVICES

SPECIAL ISSUE ON THREE-DIMENSIONAL DEVICE STRUCTURES. *IEEE Transactions on Electron Devices,* Vol. ED-25, No. 10; October, 1978.

SPECIAL ISSUE ON SOLID-STATE SENSORS, ACTUATORS, AND INTERFACE ELECTRONICS. *IEEE Transactions on Electron Devices,* Vol. ED-26, No. 12, December, 1979.

SILICON AS A MECHANICAL MATERIAL. Kurt E. Petersen in *Proceedings of the IEEE,* Vol. 70, No. 5, pages 420–457; May, 1980.

SPECIAL ISSUE ON SOLID-STATE SENSORS, ACTUATORS, AND INTERFACE ELECTRONICS. *IEEE Transactions on Electron Devices,* Vol. ED-29, No. 1; January, 1982.

THE STRUCTURE OF QUARKS AND LEPTONS

THE KEY TO THE UNIVERSE: A REPORT ON THE NEW PHYSICS. Nigel Calder. British Broadcasting Corporation, 1967.

GAUGE THEORIES OF THE FORCES BETWEEN ELEMENTARY PARTICLES. Gerard 't Hooft in *Scientific American,* Vol. 242, No. 6, pages 104–138; June, 1980.

A UNIFIED THEORY OF ELEMENTARY PARTICLES AND FORCES. Howard Georgi in *Scientific American,* Vol. 242, No. 4, pages 48–63; April, 1981.

HOT SPRINGS ON THE OCEAN FLOOR

THE EVOLUTION OF SEDIMENTARY ROCKS. Robert M. Garrels and Fred T. Mackenzie. W. W. Norton and Co., Inc., 1971.

SEAWATER-BASALT INTERACTION AT 200° C AND 500 BARS: IMPLICATIONS FOR ORIGIN OF SEA-FLOOR HEAVY-METAL DEPOSITS AND REGULATION OF SEAWATER CHEMISTRY. James L. Bischoff and Frank W. Dickson in *Earth and Planetary Science Letters,* Vol. 25, No. 3, pages 385–397; April, 1975.

THE MECHANISMS OF HEAT TRANSFER THROUGH THE FLOOR OF THE INDIAN OCEAN. Roger N. Anderson, Marcus G. Langseth and John G. Sclater in *Journal of Geophysical Research,* Vol. 82, No. 23, pages 3391–3409; August 10, 1977.

CHEMISTRY OF HOT SPRINGS ON THE EAST PACIFIC RISE AND THEIR EFFLUENT DISPERSAL. J. M. Edmond, K. L. Von Damm, R. E. McDuff and C. I. Measures in *Nature,* Vol. 297, No. 5863, pages 187–191; May 20, 1982.

A WINDOW ON THE SLEEPING BRAIN

THE FUNCTIONS OF SLEEP. Edited by René Drucker-Colin, Mario Shkurovich and M. B. Sterman. Academic Press, Inc., 1979.

SLEEP DISORDERS: INTERSECTIONS OF BASIC AND CLINICAL RESEARCH. Edited by M. H. Chase. Spectrum Publications, Inc., 1982.

DIFFERENT BEHAVIORS DURING PARADOXICAL SLEEP WITHOUT ATONIA DEPEND ON PONTINE LESION SITE. J. C. Hendricks, A. R. Morrison and G. L. Mann in *Brain Research,* Vol. 239, pages 81–105; 1982.

EARLY FARMERS OF THE NORTH EUROPEAN PLAIN

EUROPEAN PREHISTORY. Sarunas Milisauskas. Academic Press, 1978.

THE PREHISTORY OF EUROPE. Patricia Phillips. Indiana University Press, 1980.

EARLY NEOLITHIC SITES AT BRZEŚĆ KUJAWSKI, POLAND: PRELIMINARY REPORT ON THE 1976–1979 EXCAVATIONS. Ryszard Grygiel and Peter Bogucki in *Journal of Field Archaeology,* Vol. 8, No. 1, pages 9–27; 1981.

EARLY EUROPEAN AGRICULTURE. M. R. Jarman, G. N. Bailey and H. N. Jarman. Cambridge University Press, 1982.

EARLY NEOLITHIC SUBSISTENCE AND SETTLEMENT IN THE POLISH LOWLANDS. Peter I. Bogucki. British Archaeological Reports, Oxford, 1982.

CHEMICAL SIGNALS OF SOCIAL AMOEBAE

THE CELLULAR SLIME MOLDS. John Tyler Bonner. Princeton University Press, 1967.

THE DEVELOPMENT OF *DICTYOSTELIUM DISCOIDEUM.* Edited by W. F. Loomis. Academic Press, 1982.

CHEMICAL IDENTITY OF THE ACRASIN OF THE CELLULAR SLIME MOLD, *POLYSPHONDYLIUM VIOLACEUM.* O. Shimomura, H. L. B. Suthers and J. T. Bonner in *Proceedings of the National Academy of Sciences of the United States of America,* Vol. 79, No. 23, pages 7367–7379; December, 1982.

INTUITIVE PHYSICS

MEDIEVAL AND EARLY MODERN SCIENCE: VOL. 2. Alistair Crombie. Doubleday & Company, Inc., 1959.

CURVILINEAR MOTION IN THE ABSENCE OF EXTERNAL FORCES: NAÏVE BELIEFS ABOUT THE MOTION OF OBJECTS. Michael McCloskey, Alfonso Caramazza and Bert Green in *Science,* Vol. 210, No. 4474, pages 1139–1141; December 5, 1980.

STUDENTS' PRECONCEPTIONS IN INTRODUCTORY MECHANICS. John Clement in *American Journal of Physics,* Vol. 50, No. 1, pages 66–71; January, 1982.

THE AMATEUR SCIENTIST

OBSERVATIONS OF AN EARLY MORNING CUP OF COFFEE. Vincent J. Schaefer in *American Scientist,* Vol. 59, No. 5; September–October, 1971.

GRAIN SHAPE EFFECTS ON SETTLING RATES. P. D. Komar and C. E. Reimers in *Journal of Geology,* Vol. 86, No. 2, pages 193–209; March, 1978.

Marygrove College Library
Detroit, Michigan 48221
PLEASE DO NOT REMOVE

SCIENTIFIC AMERICAN

PORK PRODUCTION

$2.50

May 1983

05
74820 08715

How Mercedes-Benz harnesses the wind.

This 300SD measures almost 17 ft. long, weighs 3,780 lbs., affords 108.8 cu. ft. of interior volume. But the wedge-shaped body 1 is so skillfully contoured, that its 0.36 Cd is bettered by no other 4-door sedan sold in the U.S. today.

Brake-generated heat buildup is dissipated into airstream 2 via heavily slotted alloy wheels.

Inlet at windshield base 3 diverts measured airflow to help cool sensitive electronic components under the hood.

Channels integrated with windshield posts 4 intercept rainwater dispersed by wipers, conduct it upward onto roof area 5 and away from side windows.

Outside mirror shell 6 shaped to deflect swirling water, slush from mirror surface.

Fine aerodynamic detailing of cabin exterior 7 keeps wind turbulence and noise at minimum.

Channels set into roof 8 help guide rainwater rearward and away from side window area.

Grooves trap streaming rain and duct it around rear window frame 9 to exit at rear of car.

Ventilation gills 10, placed in a controlled low-pressure area at rear quarter panel, efficiently draw stale air from cabin. Air in cabin is completely exchanged every 20 seconds.

"Washboard" ribbing of taillights 11 helps keep their lenses clean, even in foul-weather driving.

Established 1845 **SCIENTIFIC AMERICAN** May 1983 Volume 248 Number 5

ARTICLES

53 **SMART WEAPONS IN NAVAL WARFARE, by Paul F. Walker**
Precision-guided weapons call into question plans to add costly surface ships to the U.S. fleet.

62 **VIBRATIONS OF THE ATOMIC NUCLEUS, by George F. Bertsch**
The nucleus can quiver, ring or even "breathe"; six modes of vibration have been detected so far.

74 **MICROBODIES IN THE LIVING CELL, by Christian de Duve**
They are organelles that look alike but are different. Each incorporates its own array of enzymes.

96 **MODERN PORK PRODUCTION, by Wilson G. Pond**
The interaction of economics and the biology of the pig has created a new kind of husbandry.

104 **NEW INORGANIC MATERIALS, by J. D. Birchall and Anthony Kelly**
Novel materials made out of sand, clay and other minerals require comparatively little energy.

116 **COMPUTER-INTENSIVE METHODS IN STATISTICS, by Persi Diaconis and Bradley Efron** They replace assumptions about unknown data with prodigious calculations.

132 **THE SOCIAL INFLUENCE OF THE MOTTE-AND-BAILEY CASTLE, by Michel Bur**
The appearance of this type of fortification late in the 10th century diminished central authority.

144 **SUDDEN CARDIAC DEATH: A PROBLEM IN TOPOLOGY, by Arthur T. Winfree**
The cause of fibrillation may lie in a sequence of events described by a mathematical theorem.

DEPARTMENTS

7 LETTERS
10 50 AND 100 YEARS AGO
15 THE AUTHORS
16 METAMAGICAL THEMAS
45 BOOKS
88 SCIENCE AND THE CITIZEN
162 THE AMATEUR SCIENTIST
170 BIBLIOGRAPHY

BOARD OF EDITORS Gerard Piel (Publisher), Dennis Flanagan (Editor), Brian P. Hayes (Associate Editor), Philip Morrison (Book Editor), John M. Benditt, Peter G. Brown, Michael Feirtag, Diana Lutz, Jonathan B. Piel, John Purcell, James T. Rogers, Armand Schwab, Jr., Joseph Wisnovsky

ART DEPARTMENT Samuel L. Howard (Art Director), Steven R. Black (Assistant Art Director), Ilil Arbel, Edward Bell

PRODUCTION DEPARTMENT Richard Sasso (Production Manager), Carol Hansen and Leo J. Petruzzi (Assistants to the Production Manager), Carol Eisler (Senior Production Associate), Karen O'Connor (Assistant Production Manager), Carol Albert, Lori Mogol, Martin O. K. Paul, Julio E. Xavier

COPY DEPARTMENT Sally Porter Jenks (Copy Chief), Debra Q. Bennett, Mary Knight, Dorothy R. Patterson

GENERAL MANAGER George S. Conn

ADVERTISING DIRECTOR C. John Kirby

CIRCULATION MANAGER William H. Yokel

SECRETARY Arlene Wright

SCIENTIFIC AMERICAN (ISSN 0036-8733), PUBLISHED MONTHLY BY SCIENTIFIC AMERICAN, INC., 415 MADISON AVENUE, NEW YORK, N.Y. 10017. COPYRIGHT © 1983 BY SCIENTIFIC AMERICAN, INC. ALL RIGHTS RESERVED. PRINTED IN THE U.S.A. NO PART OF THIS ISSUE MAY BE REPRODUCED BY ANY MECHANICAL, PHOTOGRAPHIC OR ELECTRONIC PROCESS, OR IN THE FORM OF A PHONOGRAPHIC RECORDING, NOR MAY IT BE STORED IN A RETRIEVAL SYSTEM, TRANSMITTED OR OTHERWISE COPIED FOR PUBLIC OR PRIVATE USE WITHOUT WRITTEN PERMISSION OF THE PUBLISHER. SECOND-CLASS POSTAGE PAID AT NEW YORK, N.Y., AND AT ADDITIONAL MAILING OFFICES. AUTHORIZED AS SECOND-CLASS MAIL BY THE POST OFFICE DEPARTMENT, OTTAWA, CANADA, AND FOR PAYMENT OF POSTAGE IN CASH. SUBSCRIPTION RATE: $21 PER YEAR IN THE U.S. AND ITS POSSESSIONS; $27 PER YEAR IN ALL OTHER COUNTRIES. POSTMASTER: SEND ADDRESS CHANGES TO SCIENTIFIC AMERICAN, 415 MADISON AVENUE, NEW YORK, N.Y. 10017.

SCIENTIFIC AMERICAN

CORRESPONDENCE

Offprints of more than 1,000 selected articles from earlier issues of this magazine, listed in an annual catalogue, are available at 75 cents each. Correspondence, orders and requests for the catalogue should be addressed to W. H. Freeman and Company, 4419 West 1980 South, Salt Lake City, UT 84121. Offprints adopted for classroom use may be ordered direct or through a college bookstore. Sets of 10 or more Offprints are collated by the publisher and are delivered as sets to bookstores.

Photocopying rights are hereby granted by Scientific American, Inc., to libraries and others registered with the Copyright Clearance Center (CCC) to photocopy articles in this issue of SCIENTIFIC AMERICAN for the flat fee of $2 per copy of each article or any part thereof. Such clearance does not extend to the photocopying of articles for promotion or other commercial purposes. Correspondence and payment should be addressed to Copyright Clearance Center, Inc., 21 Congress Street, Salem, MA 01970. Specify CCC Reference Number ISSN 0036–8733/83. $2.00 + 0.00.

Editorial correspondence should be addressed to The Editors, SCIENTIFIC AMERICAN, 415 Madison Avenue, New York, NY 10017. Manuscripts are submitted at the authors' risk and will not be returned unless they are accompanied by postage.

Advertising correspondence should be addressed to C. John Kirby, Advertising Director, SCIENTIFIC AMERICAN, 415 Madison Avenue, New York, NY 10017.

Subscription correspondence should be addressed to Subscription Manager, SCIENTIFIC AMERICAN, P.O. Box 5969, New York, NY 10017. The date of the last issue on your subscription is shown in the upper right-hand corner of each month's mailing label. For change of address notify us at least four weeks in advance. Please send your old address (if convenient, on a mailing label of a recent issue) as well as the new one.

Name

New Address

Street

City

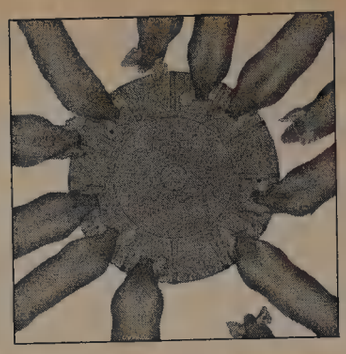
State and ZIP

Old Address

Street

City

State and ZIP



THE COVER

The painting on the cover shows a rotary feeder for pigs (see "Modern Pork Production," by Wilson G. Pond, page 96). The feeder is one of several in a "finishing operation," that is, a farm that buys pigs when they weigh from 30 to 50 pounds and feeds them until they reach the marketing weight of 220 pounds or more. The feed is a mixture of cornmeal and soybean meal, with supplements of protein, minerals and vitamins, all precisely determined to meet the nutritional needs of pigs of a certain size. On this farm pigs of the same size are fed together regardless of their age. The feed mixture is delivered to the rotary feeders by a conveyor system running overhead. The notches in the ears of the pigs tell where the animals were born; the colored ear tags indicate the auction at which each pig was bought. A strong trend in modern pig farming is toward "farrow to finish" operations, where the pig stays from birth to market.

THE ILLUSTRATIONS

Cover painting by Marvin Mattelson

Page	Source
16–19	Ilil Arbel
53	United Press International
54–57	Walken Graphics
58–59	General Dynamics Corporation
60–61	Los Alamos National Laboratory
63–73	Jerome Kuhl
75	Marten Veenhuis, State University at Groningen
76	Helen Shio, Rockefeller University (*top*); Eldon H. Newcomb, University of Wisconsin at Madison (*bottom*)
77	Eldon H. Newcomb, University of Wisconsin at Madison
78–83	Alan D. Iselin
84	Helen Shio, Rockefeller University (*top*); Pierre Baudhuin, International Institute of Cellular and Molecular Pathology (*bottom*)
97–98	Tom Prentiss
99	Ilil Arbel
100–102	Tom Prentiss
103	Ilil Arbel
105	R. F. Bonifield
106–108	Phillip A. Harrington
110–114	Ilil Arbel
115	Phillip A. Harrington (*top*), Imperial Chemical Industries (*bottom*)
117–125	Allen Beechel
126–128	Ilil Arbel
133	Michel Bur, University of Nancy
134–135	Michael Holford
136–138	Laszlo Kubinyi
145–160	George V. Kelvin
163	R. F. Bonifield
164–168	Michael Goodman
169	Jearl Walker

The air in this room must be 1,000 times cleaner than that in an operating room.

What kind of work demands such an immaculate environment?

The development of specialized, highly sophisticated microchips—done here at the Advanced Technology Center in Shelton, Connecticut, by engineers from ITT.

A microchip is, in essence, a tiny electronic brain, etched into a piece of silicon half the size of a fingernail.

The circuits on this kind of chip are so fine (a fraction as thick as a human hair) that a single speck of dust can damage a chip beyond repair.

At Shelton, the microchips made with such care and precision are prototype models, being tested for future use in ITT telecommunications systems and other applications.

Eventually, this research will produce new, more advanced ways for people to communicate with each other.

Helping people communicate has always been one of the main ways we've helped people.

And that was long before we started coming up with some of our best ideas in Shelton.

The best ideas are the ideas that help people. ITT

© 1983 International Telephone and Telegraph Corporation, 320 Park Avenue, New York, N.Y. 10022.

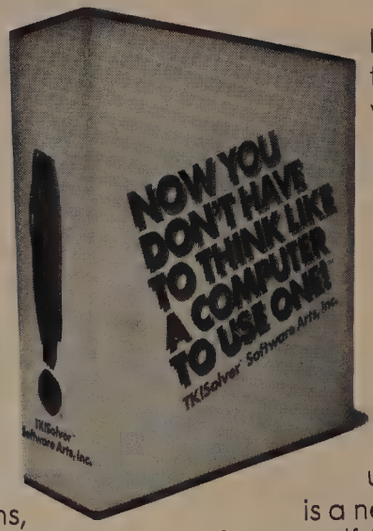
CRUNCH!

The inventor of VisiCalc,® the world's leading software product, introduces TK!Solver™—the revolutionary way to crunch problems on a personal computer.

With VisiCalc, Software Arts™ created an entirely new application for personal computers—the electronic spreadsheet.

Now Software Arts has developed TK!Solver, another entirely new and different application for personal computers—the problem solver. Business and financial problems. Scientific and engineering problems. Architectural and building design problems. Any kind of problem you can dream up, in any area.

What's so revolutionary about that? Simply this: the power of the TK!Solver program comes from the ease with which you can set up problems, vary assumptions, find solutions, and display results. All the facilities needed to solve the problems—simple or complex—are built-in, and need not be developed. Once you have stated your problem, simply enter the known values, then solve the problem with a single keystroke.



For even easier problem solving use a TK!SolverPack™ in which models for solving common problems in such areas as finance, engineering and education have already been formulated by professionals in these fields.

There is no software product in existence that solves problems as quickly, flexibly, and easily as TK!Solver—The Problem Cruncher™. In fact, no other software product even works the way TK!Solver does. It is a necessity for personal computer owners. See for yourself. Visit your nearest computer store for a demonstration and find out how much you can do with a personal computer when it thinks as flexibly as you do.

Software Arts™

27 Mica Lane, Wellesley, MA 02181

TK!Solver, TK!SolverPack and TK!Solver—The Problem Cruncher are trademarks of Software Arts, Inc. Software Arts is a trademark of Software Arts, Inc. and Software Arts Products Corp. VisiCalc is a trademark of VisiCorp. Copyright © 1983. Software Arts, Inc. All rights reserved.

LETTERS

Sirs:

I read with interest the article titled "Orphans in the Marketplace" ["Science and the Citizen," SCIENTIFIC AMERICAN, January].

I should like to add another orphan disease and drug to your list—this time a success story in that the drug has reached the marketplace. The disease is erythropoietic protoporphyria (EPP), an inherited defect of porphyrin metabolism, in which protoporphyrin accumulates in the patient's blood and skin. The accumulation causes itching and burning sensations of the skin when the patient is exposed to sunlight, in some cases making even brief excursions outdoors during the day extremely distressful, if not impossible. I hypothesized, on the basis of work done by others as well as by me on the function of carotenoid pigments in bacteria, that perhaps the administration of such pigments to patients with light-sensitivity skin diseases might prevent the annoying symptoms associated with these diseases. After successful animal testing I organized a clinical trial to test my hypothesis: we administered the carotenoid pigment beta-carotene to patients suffering from EPP. We found that in the majority of patients the administration of beta-carotene did enable them to spend significantly longer times outdoors before developing the symptoms of light sensitivity than they could before they took beta-carotene. Our findings have been confirmed by other investigators. Hoffmann–La Roche, Inc., which had provided gratis the beta-carotene for our study, then went through the laborious process of having the pigment approved as a new drug by the Food and Drug Administration. Thanks to Hoffmann–La Roche's efforts and the cooperation of the FDA, approval was obtained, and beta-carotene, marketed as Solatene, is now available for physicians to prescribe for the relief of the symptoms of light sensitivity associated with EPP. Judging by the letters I have received from EPP patients who have benefited from the use of Solatene, I can definitely say that the efforts of both the pharmaceutical company and the Government agency are much appreciated.

MICHELINE M. MATHEWS-ROTH, M.D.

Harvard Medical School
Boston, Mass.

Sirs:

Walter C. Leedy, Jr., in his interesting article on the ceilings of certain Gothic cathedrals, expressed in one sentence two popular misconceptions about igloos ["Fan Vaulting," by Walter C. Leedy, Jr.; SCIENTIFIC AMERICAN, February]. He wrote: "The carved stones were then *fitted* together to form the vaulting much as blocks of *ice* are fitted to form an igloo" (italics mine).

Most masonry is constructed so that the boundary between one layer of blocks (called a course) and the next is horizontal. The diagrams of cathedrals in the article show concentric rings of blocks piled on each other in horizontal courses. Some independent method of support such as scaffolding would have been required until the keystone was installed. Scaffolding would not be available to the lone hunter building an igloo in a windy environment.

Virtually every artistic rendering of igloos shows the line between courses to be exactly horizontal. If an igloo were built that way, one horizontal course would be completed, then a new course would have to be started. The hunter by himself would have difficulty supporting the first block of a new course while lifting the second block. This problem would be insurmountable near the top of the igloo dome, where the blocks slope inward.

The Inuit of the Canadian Arctic start an igloo by laying a complete circle of blocks on the snow. These blocks are about three feet long, two feet high and four inches thick. Then about a fourth of the blocks are trimmed down to form a ramp. A notch is created between the sloping ramp and the vertical face of the adjoining untrimmed block. The builder sets the next block into that notch. Since the blocks are roughly rectangular, a new notch is automatically formed for the next block.

Viewed from above, an igloo looks like a spiral; viewed from the side, the lines between courses slope slightly upward with respect to the snow surface (which is seldom horizontal).

When a block has been fitted, it must be firmly tapped once. A surprisingly strong bond is then developed between the block and the rest of the structure. A block will remain in place even in strong winds. I think that much of the strength of an igloo comes from this bond. A completed igloo of blocks four inches thick is strong enough to stand on. One can crawl up the side of a partially completed one.

Ice has several drawbacks as a material for igloos, even though ice blocks would be stronger than snow blocks. Ice would be much harder to shape than snow. Also the essential bond between blocks would not form. Ice blocks could slip out easily unless water were poured into the cracks. Liquid water is scarce at 40 below.

RONALD A. MACNAUGHTON, P.ENG.

Kettleby, Ontario

How neutron scattering entirely new polymer

Exxon Chemical's David Lohse can observe the structure and behavior of polymer molecules.

From easy-care fabrics to the nose cones of space vehicles, the long chain molecular structures called polymers have become, both literally and figuratively, woven into the fabric of 20th century life. Now, Exxon's David Lohse is exploring the characteristics of polymers to extend the potential of these remarkable materials.

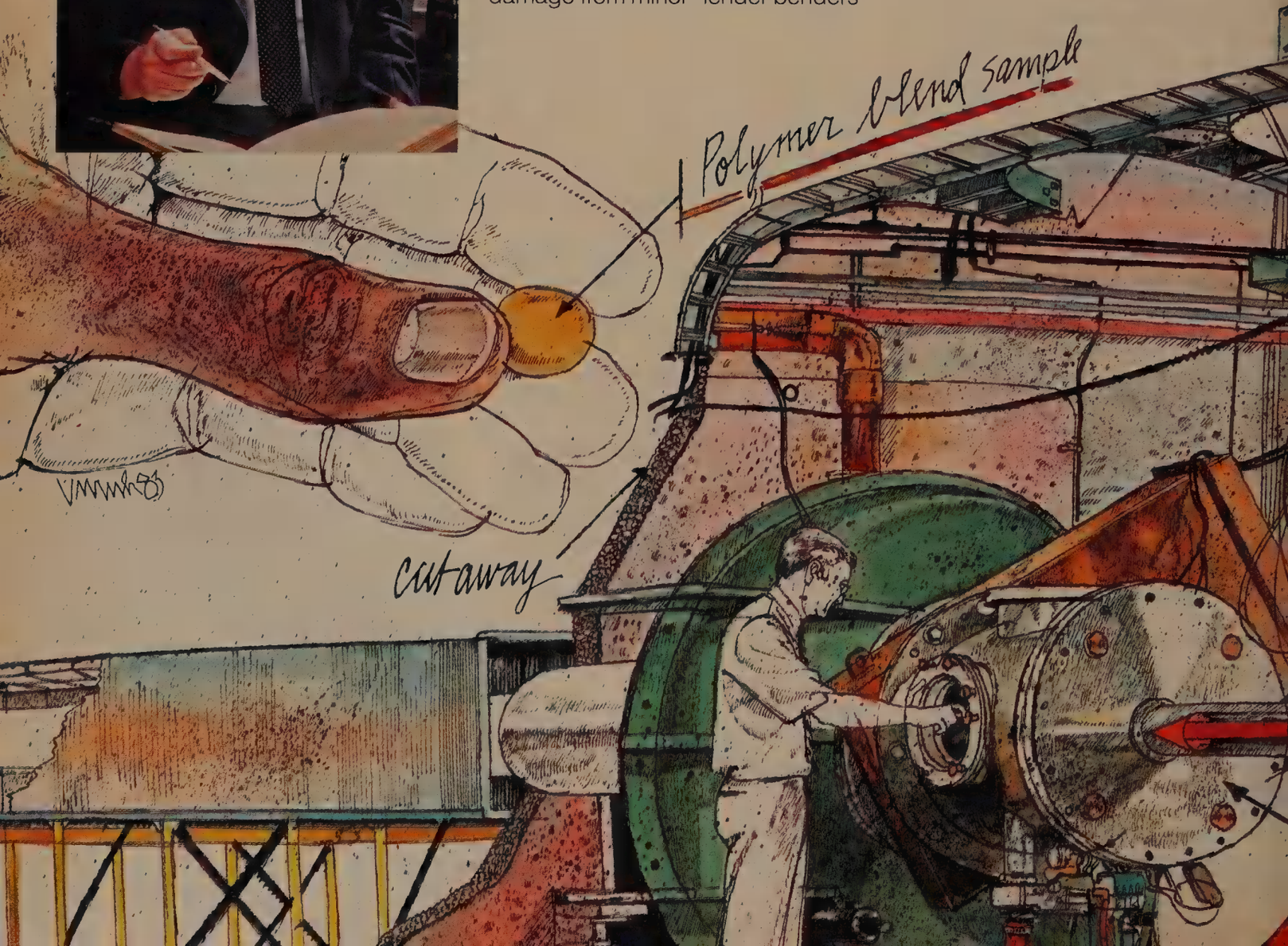
Polymer Blends

Combinations of polymer molecules can produce materials with better properties than those of the individual components. Exxon Chemical developed and is currently marketing, for example, a blend of plastic and synthetic rubber polymer material which has the formability of plastic and the elasticity of rubber. Used in automobile bumpers, this material helps reduce damage from minor "fender-benders" by absorbing impact and then retur to its original shape.

In order to develop more sophisticated blends, David Lohse and othe in Exxon Chemical's Long Range Po mer Research Group are developing the use of Small Angle Neutron Scat ing (SANS), a technique that subjec these molecules to detailed scrutiny

Small Angle Neutron Scattering

Conventional electron microscopy techniques allow scientists to study structure of a polymer blend at room temperature. But the technique, use by Exxon at the National Center for Small Angle Scattering Research (O Ridge National Laboratory) and at th National Bureau of Standards in Mar



s helping Exxon create naterials.

permits scientists to study poly-olecules at temperatures of or higher. This is crucial, since it es high temperatures for the poly- melt and be processed. In this tate, the molecules are most e and the structure of the blends ned.

Seeing Polymer Structure by SANS

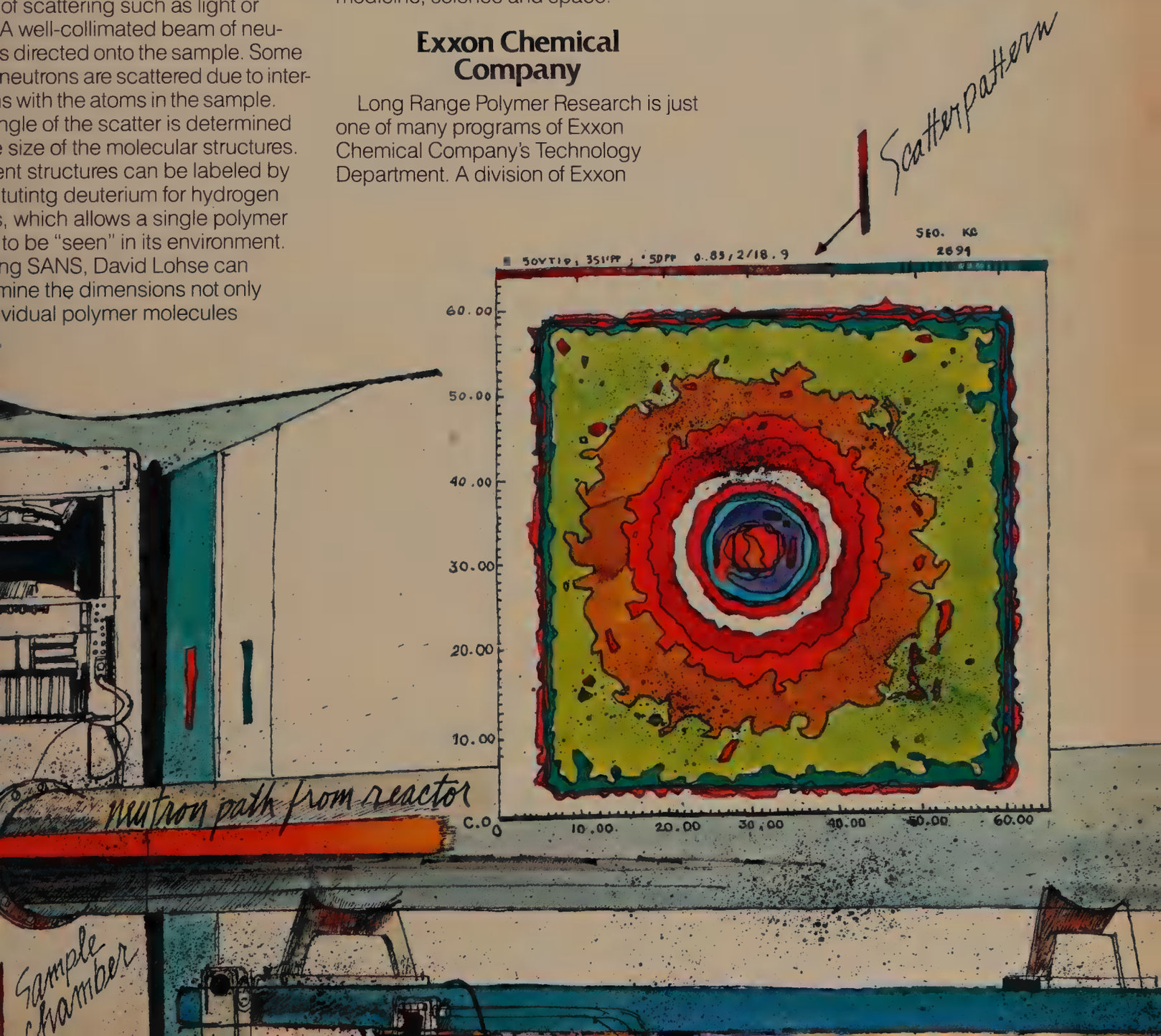
all Angle Neutron Scattering in- the same basic physics as other of scattering such as light or A well-collimated beam of neu- s directed onto the sample. Some neutrons are scattered due to inter- s with the atoms in the sample. ngle of the scatter is determined e size of the molecular structures. ent structures can be labeled by tutintg deuterium for hydrogen , which allows a single polymer to be "seen" in its environment. ng SANS, David Lohse can mine the dimensions not only vidual polymer molecules (10-100 angstroms) but also the dimensions of the phase domains of the blends (greater than 1000 angstroms). More importantly, while the blends are being heated, he can measure changes in the sizes of the individual molecules and the domains. It is these changes which determine the physical properties of the materials.

The information obtained using SANS will be used to develop new polymer blends that may find important commercial applications in the home, in industry, medicine, science and space.

Exxon Chemical Company

Long Range Polymer Research is just one of many programs of Exxon Chemical Company's Technology Department. A division of Exxon Corporation, Exxon Chemical is the world's ninth largest chemical company. Its research activities are diverse, with 2,000 scientists, researchers and technicians located in 10 countries working in advanced fields of catalysis and polymer science, and on ways to produce chemicals from new feedstocks or from more efficient processes.

EXXON



50 AND 100 YEARS AGO

SCIENTIFIC AMERICAN

MAY, 1933: "A unit volume of water has long been regarded as having a definite weight, but E. W. Washburn of the National Bureau of Standards has demonstrated very clearly that water can be as heavy as you make it. Two flasks of pure water, each containing exactly the same volume, were exhibited on a delicate balance, proving very conclusively that Dr. Washburn's water was heavier than any other pure water. This phenomenon is due to the fact that the heavy water was made from heavy hydrogen—or, in chemical parlance, from the isotope of hydrogen that has an atomic mass about twice that of ordinary hydrogen. In nature one out of about 30,000 atoms of hydrogen is double weight. When water is dissociated into hydrogen and oxygen by passing an electric current through it, the hydrogen given off contains more of the lighter hydrogen than normal. When this light hydrogen is burned, it yields a water of lower specific gravity than the water of nature. The water left in the electrolytic cell contains increasing amounts of heavy hydrogen as the electrolysis continues. In this way water that is one part in 1,000 heavier than ordinary water was obtained."

"On any cloudless, moonless spring evening, as the long western twilight fades, the sky remains brighter in one region than the rest. Rising from the point of sunset is a tapering wedge of faint light. This 'zodiacal light' has been known from ancient times. With good skies it can be seen to run on until its morning and evening halves meet or are joined by a faint 'zodiacal band.' Upon this band, and directly opposite the sun in the heavens, is a brighter spot called the Gegenschein, or zodiacal counterglow. Few if any measures of its brightness have ever been made. This gap in our observational data has recently been filled by C. T. Elvey of the Yerkes Observatory, working with a very delicate photoelectric photometer. The total brightness of the Gegenschein is surprisingly great. Elvey's measures on three nights give an average equal to the light of Arcturus. It is only because the light is so widely spread out and fades away so gradually that we find it difficult to see. The Gegenschein, the zodiacal band and the zodiacal light find a simple explanation in the assumption that a lens-shaped swarm of particles surrounds the sun and extends beyond the earth's orbit, but not very far."

"When Lindbergh flew to Japan, some newspaper comments had it that he was not flying the shortest route from the West Coast to Japan. He flew across Canada and via Alaska and the Aleutian Islands. Lindbergh was perfectly right. Our notions of distances are often based on maps in which distances are distorted. If we look down on a globe from above the North Pole, we see that the shortest way from the United States to China is over the Arctic region. The shortest way from San Francisco to England is not via New York and the North Atlantic but across Canada, Greenland, Iceland and the Faroe Islands. We must revise our notions of the shortest routes to other continents. When the United States is connected by regular airlines to China, Japan, Siberia and Europe, it may well be that these airlines will all go near the Pole. Mankind is never daunted by difficulties if a worthwhile objective is to be obtained, and plane designers, inventors and airplane flyers and operators may eventually transform the wild regions of the North into a busy sea of aerial activity."

JUNE, 1883: "The liquefaction of oxygen gas and nitrogen, and the freezing of alcohol and sulphide of carbon, are the latest achievements of chemical science. This news comes to us from the laboratory of M. Wroblewski in Cracow, Poland. By the use of liquefied ethylene M. Wroblewski and K. Olszewski obtained the remarkably low temperature of −136° C., equal to −212.8° F. Oxygen gas subjected to about this temperature and compressed to about 25 atmospheres, or 375 pounds to the square inch, was readily liquefied in glass tubes and formed a colorless and transparent liquid, very mobile and resembling carbonic acid. Nitrogen was also liquefied, forming a colorless liquid. Alcohol was solidified at −130.5° C., or −202.9° F., forming a white body. Sulphide of carbon froze at about −116° C., or −176.8° F. The difficulty heretofore experienced in the liquefaction of oxygen and nitrogen has been to obtain a sufficiently low temperature in conjunction with compression. This obstacle now appears to be removed, and a variety of new and valuable observations concerning the nature of gaseous substances can be expected."

"The underground electric railway in London, whose construction has been authorized by act of Parliament, will commence near the north end of Northumberland Avenue opposite the Grand Hotel and pass under that avenue and the Victoria Embankment to a tunnel under the Thames, thence by College Street and Vine Street to Waterloo Station, where it will form a connection with the platforms of the London and Southwestern Railway. A separate approach to the Waterloo terminus of the line will, however, be built at York Road. The line will be double and will be worked by a stationary engine at Waterloo. The cars will run singly and will start as soon as they are filled, like omnibus cars. The journey will occupy about three and a half minutes. A contract for the supply of the electrical plant has been entered into with Messrs. Siemens Brothers, and a tender for the construction of the permanent way in 18 months' time has also been accepted."

"There are several kinds of torpedoes. The one that is most used in the French navy is called the 'carried' torpedo (*torpille portée*), so named because the torpedo boat literally carries it right under the sides of the enemy's ship. It consists of a cartridge of about 20 kilogrammes of gun cotton, placed at the extremity of an iron rod, 12 meters in length, projecting in a downward direction from the fore part of the boat. The charge is fired by an electric spark by means of an apparatus placed in the lookout compartment. The water spout produced by the explosion sometimes completely covers the torpedo boat, and the latter would be sunk if it were not that all apertures were closed so as to make her a true buoy. The attack on an ironclad is a perilous operation, and one that requires men of coolness, courage and experience."

"The age in which we live is pre-eminently regardful of the rights of man. Constitutions are framed and statutes enacted for the protection of the weak against the possible encroachments of the strong. In an age and a country such as ours the very weakness of mental disease is its safeguard. A major right of the insane is the right to medical inquiry, in lieu of the ordinary trial by jury, into the question of their insanity, before committing them to asylum care and custody—such a thorough, unimpassioned medical inquiry as would certainly reach the true nature and needs of the malady, and such an inquiry is best secured by men competent from experience to investigate the nature of mental disease. An incidental right of the insane is to have proper instruction, in regard to insanity, provided for in the medical schools, and we make this demand for them: that henceforth no medical college shall be chartered that does not provide a chair of psychiatry. The true friends of the insane are in the medical profession, and its members should understand them."

CIMARRON RESPONDS

You can sense it the moment you first see Cimarron's bold, contemporary road stance. You can feel it the second you slip behind the wheel and into Cimarron's leather-faced lumbar-supported front bucket seats. And you can experience it every time you take to the open road.

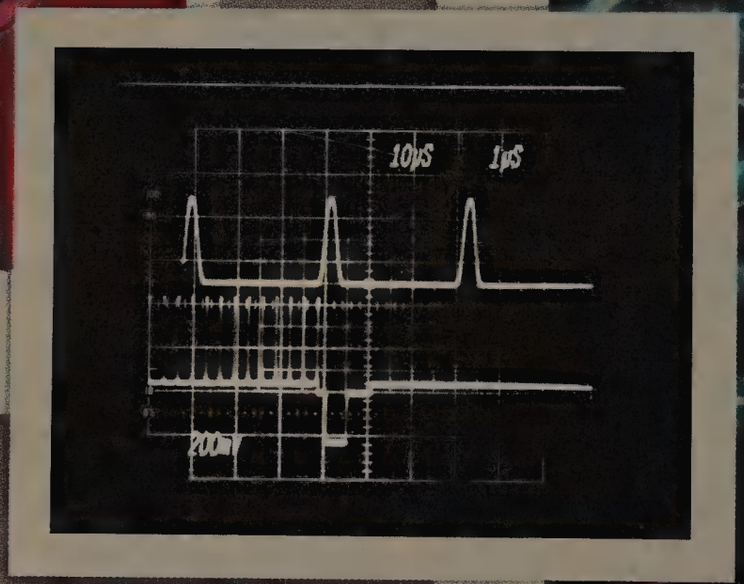
Cimarron responds. See for yourself at your Cadillac dealer's soon. And be sure to ask about the special edition Cimarron D'ORO shown above.

BEST OF ALL...IT'S A CADILLAC. CADILLAC MOTOR CAR DIVISION U.S.A.

GM

Instant te

Experimental polymer on heating stage at 310°C, polarized light, (65X) on Time-Zero Type 778 film.



Oscilloscope trace of 10 nanosecond event captured on Ultra High Speed Type 612 film.

Landsat weather image on Polacolor ER Type 809 8×10″ film. Recorded by: GOES-West Satellite. Courtesy: International Imaging Systems.

With Polaroid Instant Photography, there's no limit to what you can explore. In seconds you can have color hard copies of computer-generated images that contain more information than you'll find on even the highest resolution raster-scan display. Or capture events not visible to the naked eye. Instantly. Easily. On-the-spot.

It's all possible with our 32 different Polaroid instant professional films. They range in speed up to ISO 20,000/82°. In sizes up to 8 x 10″. They have several different spectral sensitivities. And can be used in the most sophisticated laboratory instruments, as well as in

chnology.

CAD/CAM solid modeling image of disc brake on Polacolor ER Type 559 film. Courtesy: Gray Lorig/Raster Technologies, Inc.

35mm macrophotograph and 3 1/4 × 4 1/4" Polacolor ER Type 669 print of read–write heads.

nple, hand-held cameras. No matter what you're exploring, chances are you can record it.

Polaroid instant photography gives you sharp, detailed photographic records in nsparencies or prints. You don't have to wait for processing, pay a premium for rush rvice, employ photographic help, or use valuable space and money for darkrooms. And u can be sure you have the image you need immediately. For documentation, reports, alysis and presentations.

On the following page, you'll find more information on how the images above were oduced. Because Polaroid photography discovery every scientist should make.

POLAROID

Instant Technology:

The images on the preceding pages show just a small segment of the range of explorations possible when Polaroid self-developing films are used. They also show only 5 of the 32 different Polaroid instant professional film types available. **1.** Color photomicrograph of polymer taken on Polaroid self-developing Time-Zero Type 778 film (ISO 150/23°) with the SX-70 Microscope Adapter, which features automatic exposure control. **2.** Instant 8×10″ color hard copy or Colorgraph overhead projection transparencies of computer graphics produced in just minutes on Polaroid VideoPrinter Model 8. **3.** Oscilloscope trace taken in just 30 seconds with hand-held Polaroid CU-5 Close Up Camera and recorded on Type 612 Ultra High Speed 3 1/4 × 4 1/4″ black and white film (ISO 20,000/82°). **4.** Polaroid 3 1/4 × 4 1/4″ color or black and white print of 35mm slide made at the touch of a button with the portable Polaprinter Slide Printer. **5.** Instant color 4×5″ or 3 1/8 × 3 1/8″ working size hard copies, and conventional or new Polaroid 35mm Autoprocess* transparencies of computer-generated images can be produced on-the-spot using Polaroid's VideoPrinter Model 4.

1.

2.

3.

4.

5.

© 1983 Polaroid Corporation "Polaroid," "Polacolor," "Colorgraph," "Time-Zero," "SX-70" and "Polaprinter"* *Scheduled for availability later this year.

Instant Response:

If you have any specific questions on the uses or performance of Polaroid films, or would like information on how to adapt your equipment to accommodate our films, call us, toll-free in the continental U.S.: 800-225-1618. (In Massachusetts, call collect: 617-547-5177.)

For more general information see your Polaroid Photographic Dealer or use the coupon.

POLAROID

Polaroid Corporation, Industrial Marketing Dept. 567, P.O. Box 5011, Clifton, NJ 07012

Please send me appropriate literature. I am interested in: _____

_____

Name: _____

Company: _____

Address: _____

City: _____

State: _____ Zip: _____

Phone: _____

SA-5/83

THE AUTHORS

PAUL F. WALKER ("Smart Weapons in Naval Warfare") is director of the division of education and programs of Physicians for Social Responsibility. After being graduated from the College of the Holy Cross he joined the U.S. Army, where he was a Russian translator and intelligence specialist, among other duties. On leaving the Army he went to graduate school, receiving his M.A. in international relations at the Johns Hopkins University School of International Relations in 1973 and his Ph.D. in political science from the Massachusetts Institute of Technology in 1978. Before taking up his current job he was director of research on arms-control issues for the Union of Concerned Scientists. Walker is one of six coauthors of *Winding Down: The Price of Defense,* published by W. H. Freeman and Company.

GEORGE F. BERTSCH ("Vibrations of the Atomic Nucleus") is professor of physics at Michigan State University. He attended Swarthmore College as an undergraduate, getting his bachelor's degree in 1962. He went on to earn his Ph.D. in physics from Princeton University in 1965. After serving on the faculties of Princeton and the Massachusetts Institute of Technology he moved to Michigan State in 1970. He writes: "Physics has interested me since I was a teenager in New York. When I discovered in graduate school that experimental physics requires long hours of work late at night, I chose theoretical physics for a career. My main research goal is to understand the cooperative behavior of many-particle systems, with nuclei being the subject of most of my work."

CHRISTIAN DE DUVE ("Microbodies in the Living Cell") is a cell biologist who divides his working time among three widely separated jobs. He is founder and president of the International Institute of Cellular and Molecular Pathology in Brussels, Andrew W. Mellon Professor at Rockefeller University and professor and chairman of the department of physiological chemistry at the Faculty of Medicine of the Catholic University of Louvain. He is a citizen of Belgium who was born in England in 1917. He got his M.D. at Louvain in 1941 and remained there to earn a master's degree in 1946. In 1949, employing a method of cell fractionation developed by Albert Claude, de Duve and his colleagues at Louvain discovered the lysosome, an organelle of the living cell. For this work, and for other work on the structure and function of the cell, de Duve (with Claude and George E. Palade) received the Nobel prize for medicine or physiology in 1974. He moved to Rockefeller University in 1962 but kept his professorship at Louvain. He founded the Institute of Cellular and Molecular Pathology in 1975.

WILSON G. POND ("Modern Pork Production") is research leader in nutrition at the Roman L. Hruska Meat Animal Research Center of the U.S. Department of Agriculture in Clay Center, Neb. He got his B.S. at the University of Minnesota in 1952. His Ph.D. in animal nutrition, granted in 1957, is from Oklahoma State University. From 1957 until 1978 he was a member of the department of animal science at Cornell University. In 1978 he went on to the Hruska Center. Pond's main scientific interest is the subject of his article. He has served as a consultant to pork producers in Japan, Italy, Germany, Taiwan and Ecuador as well as the U.S.

J. D. BIRCHALL and ANTHONY KELLY ("New Inorganic Materials") are respectively senior research associate at Imperial Chemical Industries (ICI) and vice-chancellor of the University of Surrey. Birchall writes: "I joined the research staff of ICI in 1957 following a period of working in small companies and a brief period as a self-employed inventor. My first patent was granted at the age of 19. At ICI, I became interested in crystal nucleation and growth. This interest widened to include studies of the behavior of crystalline solids on heating." Birchall also is visiting professor at the University of Surrey. Kelly was graduated from the University of Reading with a B.Sc. in 1949 and received his Ph.D. in physics from the University of Cambridge in 1953. After academic appointments at several institutions, including Northwestern University and the University of Cambridge, he moved to the National Physical Laboratory in Britain in 1967 as superintendent of the Division of Inorganic and Metallic Structure. For part of his tenure at the National Physical Laboratory he was seconded to ICI as part of a task force for encouraging exchange among business, government and universities. Kelly took up his current job in 1975.

PERSI DIACONIS and BRADLEY EFRON ("Computer-Intensive Methods in Statistics") are both professors of statistics at Stanford University. Diaconis' first profession was that of illusionist. He left home at 14 to work with Dave Vernon, a well-known expert in sleight of hand. He spent 10 years traveling as Vernon's assistant before enrolling as an undergraduate at the City College of the City University of New York. He got his B.S. there in 1971 and three years later had earned his Ph.D. in statistics from Harvard University. His research interests include Bayesian statistics and the use of probability to analyze computer algorithms. Diaconis writes that his "background in magic can be seen in papers that point out errors in modern parapsychology and in work on questions such as 'How many times do you have to shuffle a deck of cards until it is close to random?'" Efron writes: "I received my bachelor's degree in mathematics from the California Institute of Technology, came to Stanford in 1960 as a graduate student, received my Ph.D. in statistics in 1965 and have been essentially unmovable ever since. My half-time appointment in the medical school as professor of biostatistics has provided the applied impetus for much of my theoretical work."

MICHEL BUR ("The Social Influence of the Motte-and-Bailey Castle") is professor of medieval history at the University of Nancy. He was trained as a historian and wrote his doctoral dissertation on the formation of the Comté de Champagne. He writes that "because of the paucity of documentary evidence" pertaining to the Middle Ages, much of his work is now archaeological. He has two major current projects. He is director of an inventory of the fortified buildings of the Middle Ages in France, which is being done under the auspices of the Centre National de la Recherche Scientifique. He is also analyzing the results of excavations at a 13th-century *château.* Bur adds: "I am aware that archaeological research rarely becomes part of the discourse of historians, and my preoccupation is to make a bridge between archaeology and history."

ARTHUR T. WINFREE ("Sudden Cardiac Death: A Problem in Topology") is professor of biological sciences at Purdue University. He studied engineering physics as an undergraduate at Cornell University, from which he was graduated in 1965. He continued his study of biology and physics at Princeton University; his Ph.D. in biology (1970) is from Princeton. From 1969 to 1972 he was assistant professor of theoretical biology at the University of Chicago. At the end of that time he went on to Purdue. He writes that three questions guide his current work: "How can we decipher the curious regularities of human sleep-wake timing? What are the organizing centers that form and maintain themselves in excitable media? What is the turbulence called fibrillation in heart muscle and how does it start?" Winfree is currently on sabbatical leave from Purdue as a John Simon Guggenheim Memorial Fellow.

METAMAGICAL THEMAS

Computer tournaments of the Prisoner's Dilemma suggest how cooperation evolves

by Douglas R. Hofstadter

Life is filled with paradoxes and dilemmas. Sometimes it even feels as if the essence of living is the sensing—indeed the savoring—of paradox. Although all paradoxes seem somehow related, some seem abstract and philosophical whereas others touch on life directly. A very lifelike paradox is the Prisoner's Dilemma, discovered in about 1950 by Merrill M. Flood and Melvin Dresher, and later formalized by Albert W. Tucker. I shall present it first as a metaphor and then as a formal problem.

The original formulation in terms of prisoners is a little less clear to the uninitiated, in my experience, than the following one. Assume you possess large quantities of some item (money, for example) and want to obtain some amount of another item (stamps, groceries, diamonds). You arrange a mutually agreeable trade with the only dealer of that item known to you. You are both satisfied with the amounts you will be giving and getting. For some reason, though, the exchange must take place in secret. Each of you agrees to leave a bag at a designated place in the woods and to pick up the other's bag at the other's designated place. Suppose it is clear to both of you that you will never meet or have further dealings with each other again.

Clearly there is something for each of you to fear, namely that the other one will leave an empty bag. Obviously if you both leave full bags, you will both be satisfied, but equally obviously it is even more satisfying to get something for nothing. You are therefore tempted to leave an empty bag. In fact, you can even reason it through with seeming rigor this way: "If the dealer brings a full bag, I'll be better off having left an empty bag, because I'll have got all I wanted and given away nothing. If the dealer brings an empty bag, I'll be better off having left an empty bag, because I'll not have been cheated. I'll have gained nothing but lost nothing either. Thus it seems that *no matter what the dealer chooses to do* I'm better off leaving an empty bag. And so I'll leave an empty bag."

The dealer, meanwhile, being in more or less the same boat (although at the other end of it), thinks analogous thoughts and comes to the parallel conclusion that it is best to leave an empty bag. And so both of you, with your impeccable (or seemingly impeccable) logic, leave empty bags and go away empty-handed. How sad, because if you had both just cooperated, you could each have gained something you wanted to have. Does logic prevent cooperation? That is the issue presented by the Prisoner's Dilemma.

In case you are wondering why it is called the Prisoner's Dilemma, here is the reason. Imagine that you and an accomplice (someone you have no feelings for one way or the other) committed a crime; now you have both been caught and jailed and are apprehensively awaiting trials. You are being held in separate cells with no way to communicate with each other. The prosecutor offers each of you the following deal (and tells you both that the identical deal is being offered to each of you—and that you both know it): "We have a lot of circumstantial evidence on you both. So if you both claim innocence, we'll convict you anyway and you'll both get two years in jail. But if you'll help us out by admitting your guilt and making it easier for us to convict your accomplice—pardon me, your alleged accomplice—why then we'll let you out free. And don't worry about revenge—your accomplice will be in for five years. How about it?" Warily you ask, "But what if we both say we're guilty?" "Ah, well, my friend—I'm afraid then you'll both get four-year sentences."

		Dealer	
		Cooperates	Defects
You	Cooperate	(2,2)	(−1,4)
	Defect	(4,−1)	(0,0)

Where (x,y) means you get x points and the dealer gets y points.

A payoff matrix for the Prisoner's Dilemma

Now you are in a pickle! Clearly you do not want to maintain your innocence if your partner has confessed, because then you are going to jail for five long years. It is better you should both confess; then you will get only four. On the other hand, if your partner maintains innocence, then the best possible thing for you to do is to confess, since then you go scot-free. Hence at first it seems obvious what you should do: sing. What is obvious to you, however, is equally obvious to your partner, and so now it looks as if you both ought to sing, which means—Sing Sing for four years! At least that is what logic tells you to do. It's funny, since if both of you had been illogical and had maintained your innocence, you would both be in for only half as long. Ah, logic does it again.

Let us now go back to the original metaphor and slightly alter its conditions. Suppose both you and the dealer very much want to have a regular supply of what the other has to offer, and so before conducting your first exchange you agree to carry on a lifelong exchange once a month. You still expect never to meet face to face. In fact, neither of you has any idea how old the other is, so that you cannot be sure how long this lifelong agreement will last, but it seems safe to assume it will last for a few months anyway, and quite likely for years.

Now, what do you do on your first exchange? Bringing an empty bag seems fairly nasty as the opening of a relationship; hardly an effective way to build up trust. Suppose you bring a full bag and the dealer brings one too. All is bliss—for a month. Then you both must go back. Will your next bag be empty or full? Each month you have to decide whether to "defect" (bring an empty bag) or to "cooperate" (bring a full one). Suppose one month, unexpectedly, your dealer defects. Now what do you do? Will you decide that the dealer can never be trusted again and that from now on you will always bring an empty bag, in effect totally giving up on the entire project? Or will you pretend you didn't notice and continue being friendly? Or will you try to punish the dealer by some number of defections of your own? One? Two? A random number? An increasing number, depending on how many defec-

Lake Joseph, Ontario/Canada

Breezin' Along

Golden hours in our great out-of-doors, warm summer sunshine, the smell of pine, the riff and pull of the sail unfurling, and all kinds of ways to get away from it all – from the trails or the courts to the links or watersports, zestful or restful – Ontario's resorts harbour magic moments far from the madding crowd. Slip away. Call TOLL FREE 1-800-828-8585 or from New York State 1-800-462-8404 or write: Ontario Travel, Queen's Park, Toronto M7A 2E5, Ontario/Canada.

ONTARIO
yours
to discover!

THE CAR YOU CAN HOLD ON TO ONE THIRD LONGER.

If you're about to go shopping for a new car, get a grip on yourself. Prices are astronomical.

But if you buy a Volvo, you'll get more than your money's worth. Studies show Volvo's life expectancy is now one third longer than the average of all other cars on the road.*

Besides getting more years for your money, you get more car. All Volvos come with power assisted brakes and steering, fuel injection, radials, even reclining front bucket seats with adjustable lumbar supports.

So in addition to being the car you can hold on to longer, you'll find Volvo is also the car you won't want to let go of.

VOLVO
A car you can believe in.

*Analysis conducted by Kerr, Warwick & Associates, Inc. based on a comparison of U.S. registration data for Volvo and for the automobile industry as a whole (but not by individual make). This is not, however, an expressed or implied warranty, but it's comforting to know. © 1982 Volvo of America Corporation.

tions you have experienced? Just how mad will you get?

This is the Iterated Prisoner's Dilemma. It is a very difficult problem. It can be, and it has been, rendered more quantitative and in that form studied with the methods of game theory and computer simulation. How does one quantify it? One builds a "payoff matrix" presenting point values for the various alternatives. A typical matrix is shown on page 16.

In this matrix mutual cooperation earns both parties 2 points (the subjective value of getting a full bag of what you want and giving up a full bag of what you have). Mutual defection earns you both zero points (the subjective value of gaining nothing and losing nothing, apart from making a vain trip out to the woods that month). Cooperating while the dealer defects stings: you get −1 point whereas the rat gets 4 points. Why so many? Because it is so pleasurable to get something for nothing. And, of course, if *you* happen to be a rat some month when the dealer has cooperated, then you get 4 points and he gets −1.

It is obvious that in a collective sense it would be best for both of you to always cooperate. We are assuming, however, that you have no regard whatsoever for the other person. There is no "collective good" you are both working for. You are both supreme egoists. Then what? The meaning of "egoist" can perhaps be made clear by the following. Suppose you and your dealer have developed a trusting relationship of mutual cooperation over the years, but one day you receive secret and reliable information that the dealer is quite sick and will soon die, probably within a month or two. The dealer has no reason to suspect you have heard. Are you not powerfully tempted to defect all of a sudden, in spite of all your years of cooperating? You are, after all, looking out for yourself and no one else in this cruel world. And since it seems this may well be the dealer's last month, why not profit as much as possible from your secret knowledge? Your defection may never be punished, and at the worst it will be punished by one last-gasp defection by the dying dealer.

The surer you are that this next turn is to be the very last one, the more you feel you must defect. Either of you would feel that way, of course, on learning the other one was nearing the end of the line. This is what is meant by "egoism." It means you have no feeling of friendliness or good will or compassion for the other player; you have no conscience; all you care about is amassing points, more and more and more of them.

What does the payoff matrix for the other metaphor, the one involving prisoners, look like? You will find it at the top of the page.

		Your accomplice	
		Stays mum	Sings
You	Stay mum	(−2,−2)	(−5,0)
	Sing	(0,−5)	(−4,−4)

Where $(-x,-y)$ means you get x years in jail and your accomplice gets y years.

A payoff matrix for actual prisoners in the Prisoner's Dilemma

The equivalence of this matrix and the preceding one is clear if you add a constant—namely 4—to all terms in this one. Indeed, we could add any constant to either matrix and the dilemma would remain essentially unchanged. And so let us add 5 to this one to get rid of all negative "payoffs." We get the "canonical" Prisoner's Dilemma payoff matrix at the bottom of the page.

The number 3 is called the "reward for mutual cooperation," or R for short. The number 1 is called the "punishment," or P. The number 5 is T, the "temptation," and 0 is S, the "sucker's payoff." The conditions that make a matrix represent a Prisoner's Dilemma situation are these:

$$(1) \qquad T > R > P > S$$

$$(2) \qquad \frac{T+S}{2} < R$$

The first condition simply makes the argument go through for each of you that it is "better for me to defect no matter what my counterpart does." The second one simply guarantees that if the two of you somehow get locked into out-of-phase alternations (that is, "You cooperate, I defect" one month and "You defect, I cooperate" the next), you will not do better—in fact, you will do worse—than if you were cooperating each month.

Well, what would be your best strategy? It can be shown quite easily that there is no universal answer to the question. That is, there is no strategy better than all other strategies under all circumstances. Consider the case where the other player is playing ALL D: the strategy of defecting on every round. In that case the best you can possibly do is to defect every time, including the first time, yourself. On the other hand, suppose the other player is using the Massive Retaliatory Strike strategy, which means "I'll cooperate until you defect and thereafter I'll defect forever." In that case if you defect on the very first move, you will get one T and all P's thereafter until one of you dies. If you had waited to defect, you could have benefited from a relationship of mutual cooperation, amassing many R's beforehand. Clearly that bunch of R's will add up to more than the single T if the game goes on for more than a few moves. This means that against the ALL D strategy the best counterstrategy is ALL D, whereas against Massive Retaliatory Strike the best counterstrategy is "Always cooperate unless you learn that you or the other player is just about to die, in which case defect." This simple argument shows that how you should play depends on the other player.

The entire concept of the "quality" of a strategy takes on a decidedly more operational and empirical meaning if one imagines an ocean populated by dozens of little organisms swimming around and playing Prisoner's Dilemma over and over with each other. Suppose each time two such organisms encounter each other they recognize each other and remember how previous encounters have gone. This enables each one to decide what it wants to do this time. Now, if each organism is continually swimming around and bumping into the others, eventually each one will have met every other one numerous times and therefore all strategies will have been given the opportunity to interact with one another. By "interact" what is meant is certainly not that any one organism knocks any other one out of the ocean, as in an elimination tournament. The idea is simply that each organism gains zero or more points in each meeting, and if enough time is allowed to elapse, each organism will have met with every other one about the same number of times, and now the only question is: Which organism has amassed the most points?

It does not help organism X if it has "beaten" organism Y, in the sense that X gained more from interacting with Y than Y gained from interacting with X. That kind of "victory" is irrelevant here. What matters is not the number of "victories" rung up by any organism but the organism's *total point count*—a number that measures its overall viability in this particular "sea" of many strategies. It sounds nearly paradoxical, but an or-

		Player *B*	
		Cooperates	Defects
Player *A*	Cooperates	(3,3)	(0,5)
	Defects	(5,0)	(1,1)

Canonical Prisoner's Dilemma payoff matrix

ganism could lose many, indeed all, of its individual skirmishes with other organisms and still come out the overall winner.

As the image suggests, the situation is relevant to questions in evolutionary biology. Can totally selfish and unconscious organisms living in a common environment come to evolve reliable cooperative strategies? Can cooperation emerge in a world of pure egoists? In a nutshell, can cooperation evolve out of noncooperation? If it can, it has revolutionary import for the theory of evolution, because many critics of evolutionary theory have maintained that this was one place where it was hopelessly snagged.

As it happens, it has now been definitively demonstrated that such cooperation can emerge, and it was done through a computer tournament conducted by Robert Axelrod of the Department of Political Science and the Institute for Public Policy Studies at the University of Michigan at Ann Arbor. More accurately, Axelrod first studied by means of a computer tournament the ways cooperation evolved, and when general trends emerged, he was able to perceive the underlying principles and to prove theorems that established the facts and conditions of the rise of cooperation out of nothing. Axelrod has written a remarkably thought-provoking book on his findings: *The Evolution of Cooperation,* soon to be published by Basic Books. Furthermore, he and William D. Hamilton, an evolutionary biologist, have worked out and published many of the implications of these discoveries for evolutionary theory. Their work has won much notice—including the 1981 Newcomb Cleveland Prize, which is awarded annually by the American Association for the Advancement of Science for "an outstanding paper published in *Science.*"

There are really three aspects of the question "Can cooperation emerge in a world of egoists?" The first aspect is: How can cooperation get started at all? The second is: Can cooperative strategies survive better than their noncooperative rivals? The third is: Which cooperative strategies will do best, and how will they come to predominate?

To make these issues vivid let me describe Axelrod's tournament and its astonishing results. In 1979 Axelrod sent out invitations to a number of professional game theorists, including people who had published studies of the Prisoner's Dilemma, telling them that he sought to pit many strategies against one another in a round-robin Prisoner's Dilemma tournament, with the overall goal for each strategy being to amass more points than any other strategy. He asked for strategies to be encoded as computer programs that could respond to the *C* (cooperation) or *D* (defection) of another player, taking into account the remembered history of previous meetings with that same player. A program should always reply with a *C* or a *D,* of course, but its choice need not be deterministic. That is, consultation of a generator of random numbers was allowed at any point in a strategy.

Fourteen entries were submitted to Axelrod, and he introduced into the field one more program, called RANDOM, which in effect flipped a coin (computationally simulated, to be sure) on each move: cooperating if heads came up, defecting if tails came up. The field was a rather diverse one, consisting of programs ranging from as few as four lines to as many as 77 (of the computer language Basic). Every program was made to engage every other program (and a clone of itself) 200 times. The tournament was actually run five times in a row, so that pseudoeffects caused by statistical fluctuations in the random-number generator would be smoothed out by averaging.

The program that won was submitted by an old hand at the Prisoner's Dilemma: Anatol Rapoport, a psychologist and philosopher at the University of Toronto. His was the shortest of all programs submitted, and it is called TIT FOR TAT. TIT FOR TAT has a very simple tactic, which one might call a tacfortic: Cooperate on move 1; thereafter do whatever the other player did on the previous move. That is all. It sounds outrageously simple. How in the world could such a program defeat the complex stratagems devised by other experts?

Well, Axelrod maintains that the game theorists in general did not go far enough in their analysis. They looked "only two levels deep," when to do better they should have looked *three* levels deep. What precisely does this mean? Axelrod takes a specific case to illustrate his point. Consider the entry called JOSS (submitted by Johann Joss, a mathematician in Zurich). JOSS's strategy is quite similar to TIT FOR TAT's in that it begins by cooperating, always responds to defection by defecting and *nearly* always responds to cooperation by cooperating. The hitch is that JOSS uses a random-number generator to help it decide when to pull a "surprise defection" on the other player. JOSS is set up so that it has a 10 percent probability of defecting right after the other player has cooperated.

In playing with TIT FOR TAT, JOSS will do fine until it tries to catch TIT FOR TAT off guard. When JOSS defects, TIT FOR TAT retaliates with a single defection, whereas JOSS "innocently" goes back to cooperating. Thus we have a *DC* pair. On the next move the *C* and the *D* will switch places, since each program echoes the other's latest move, and so it will go: *CD,* then *DC, CD, DC* and so on. There may ensue a long reverberation triggered by JOSS's *D,* but sooner or later JOSS will randomly throw in *another* unexpected *D* after a *C* from TIT FOR TAT. At this point there will be a *DD* pair, and that determines the rest of the match. Both strategies will now defect forever. The echo effect resulting from JOSS's first attempt at exploitation and TIT FOR TAT's simple punitive act leads ultimately to complete distrust and lack of cooperation.

This may seem to imply that both strategies are at fault and will suffer for it at the hands of others, but in fact the one that suffers from it most is JOSS, since JOSS tries the same trick on partner after partner, which in many cases leads to the same type of breakdown of trust, whereas TIT FOR TAT, never defecting first, will never be the initial cause of a breakdown of trust. Axelrod's technical term for a strategy that never defects before its opponent does is "nice." TIT FOR TAT is a nice strategy, JOSS is not. It should be noted that "nice" does not mean that a strategy *never* defects. TIT FOR TAT defects when it is provoked, but that is still considered being nice.

Axelrod summarizes the first tournament: "A major lesson of this tournament is the importance of minimizing echo effects in an environment of mutual power. A sophisticated analysis must go at least three levels deep. First is the direct effect of a choice. This is easy, since a defection always earns more than a cooperation. Second are the indirect effects, taking into account that the other side may or may not punish a defection. This much was certainly appreciated by many of the entrants. But third is the fact that in responding to the defections of the other side one may be repeating or even amplifying one's own previous exploitative choice. Thus a single defection may be successful when analyzed for its direct effects, and perhaps even when its secondary effects are taken into account. But the real costs may be in the tertiary effects when one's own isolated defections turn into unending mutual recriminations. Without their realizing it, many of these rules actually wound up punishing themselves. With the other player serving as a mechanism to delay the self-punishment by a few moves, this aspect of self-punishment was not perceived by the decision rules....

"The analysis of the tournament results indicates that there is a lot to be learned about coping in an environment of mutual power. Even expert strategists from political science, sociology, economics, psychology and mathematics made the systematic errors of being too competitive for their own good, not for-

12 flights a week are 12 too many, Qantas.

Once the koala roamed The Outback unrestrained. The eucalypti were lush. Life was serene.

Then came Qantas.

And on Qantas came people. Business people. Tourist people. Too many people.

Today Qantas brings more people to Australia than any other airline in the world. Because Qantas has more flights to Australia than any other airline.

Qantas flies every day from San Francisco. Five days a week from Los Angeles. Twice a week from Vancouver.

Which means they're coming into Sydney daily. And into Melbourne daily. And all too often into Brisbane, Perth, Townsville and Darwin.

We demand Qantas cut back this unreasonable schedule. Unite with us.

Koalas Against Qantas. (The Australian Airline.)

KING FAISAL INTERNATIONAL PRIZE

General Secretariat

AN INVITATION TO THE NOMINATION FOR THE KING FAISAL INTERNATIONAL PRIZE IN SCIENCE

The General Secretariat of The King Faisal International Prize, at Riyadh, Kingdom of Saudi Arabia, has the honour to invite the Universities, Academies, Educational Institutions and Research Centres all over the World to nominate qualified candidates for The King Faisal International Prize to be awarded in Science which has been postponed to Rabi-al-Awal 1404 AH. i.e. January 1984.

(a) The topic of the prize will be "PHYSICS".

(b) Selection shall be according to the discretion and decision of a Committee consisting of National and International assessors selected by The Board of King Faisal International Prize.

(c) More than one person may share the prize.

(d) The winner's names will be announced in January 1984, and the prize will be awarded in an official ceremony to be held for that purpose in Riyadh, Kingdom of Saudi Arabia.

(e) The prize consists of :

(1) A certificate in the name of the winner containing abstract of his work that qualified him for the prize.

(2) A precious medal.

(3) A sum of two hundred fifty thousand Saudi Riyals (S.R. 250,000).

(f) Nominees should satisfy the following conditions:

1. A nominee must have accomplished an outstanding academic work in the subject of the prize leading to the benefit of mankind and enrichment of human thought.
2. The work submitted with the nomination of the prize must have already been printed and published. If possible an abstract in Arabic should be attached if the work is published in any other language.
3. The prize will be awarded for specific original research but the lifetime background of work will be taken into account.
4. The specific work submitted must not have been awarded a prize by any international educational institution, scientific organization, or foundation.
5. The nomination must be submitted by leading members of recognised educational institutions and of world-fame such as Universities, Academies and Research Centers. The nominations of other individuals and political parties will not be accepted.
6. The nominations must give full particulars of the nominee's academic background, experiences and/or his publications, copies of his educational certificates, if available, and three 6 x 9 cm photographs. The nominee's full address and telephone number are also requested.
7. The nominations and works in ten copies are to be sent by registered air mail to the address stated below.
8. The latest date for receipt of the full nominations with copies of works is the 12th of Dhu Al-Qe'dah, 1403 AH. i.e. the 20th of August 1983.
9. No nomination papers or works will be returned to the senders.
10. Enquiries should be made, and nominations should be sent, to the Secretary General of the King Faisal International Prize, P. O. Box 22476 , Riyadh 11495, Kingdom of Saudi Arabia , Telex 204667 PRIZE SJ.

giving enough and too pessimistic about the responsiveness of the other side."

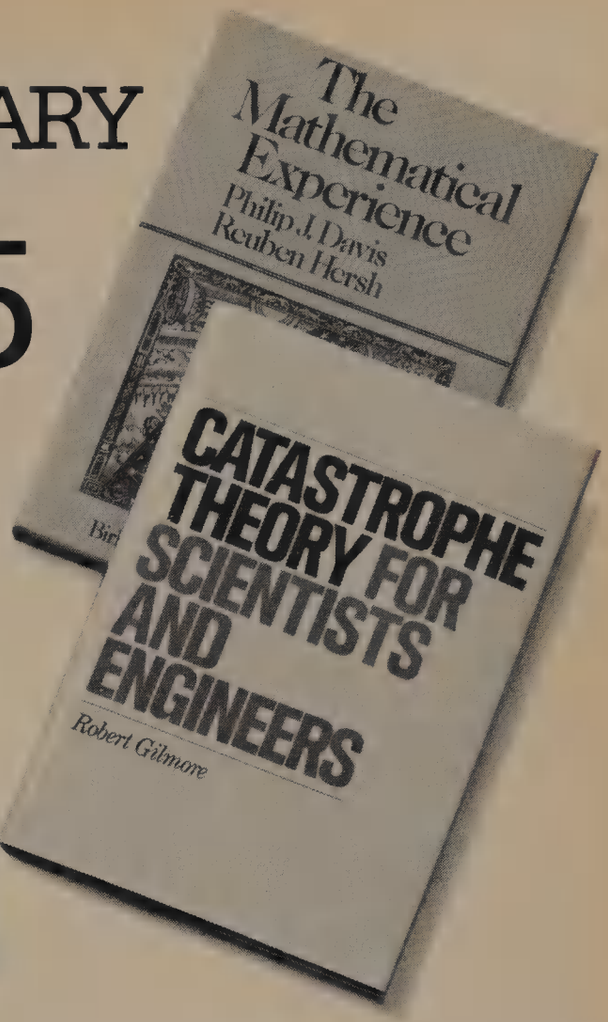
Axelrod not only analyzed the first tournament; he even performed a number of "subjunctive replays" of it, that is, replays with different sets of entries. He found, for instance, that the strategy called TIT FOR TWO TATS, which tolerates two defections before getting mad (but still only strikes back once), *would* have won, had it been in the lineup. Similarly, two other strategies he discovered, one called REVISED DOWNING and one called LOOK-AHEAD, would have come in first if they had been in the tournament.

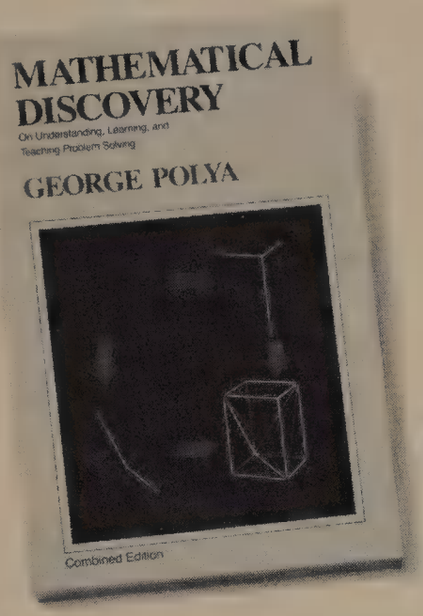
The summarized lesson of the first tournament seems to have been that it is important to be *nice* ("Do not be the first to defect"), and *forgiving* ("Do not hold a grudge once you have vented your anger"). TIT FOR TAT has both qualities.

After this careful analysis Axelrod felt that significant lessons had been unearthed, and he was convinced that more sophisticated strategies could be concocted based on the new information. Therefore he decided to hold a larger computer tournament. For this tournament he not only invited all the participants in the first round but also advertised in computer-hobbyist magazines, hoping to attract people who were addicted to programming and who would be willing to devote a good deal of time to working out and perfecting their strategy. To each person who entered Axelrod sent a detailed analysis of the first tournament, along with a discussion of the "subjunctive replays" and the strategies that would have won. He described the strategic concepts of niceness and forgiveness that seemed to capture the lessons of the tournament, and also strategic pitfalls to avoid. Naturally each entrant realized that all the other entrants had received the same information, so that everyone knew that everyone knew that everyone knew that....

There was a large response to Axelrod's call for entries. Entries were received from six countries, from people of all ages and from eight different academic disciplines. Rapoport entered again, resubmitting TIT FOR TAT (and was the only one to submit it, even though it was explicitly stated that people could enter any program they wanted). A 10-year-old entered, as did one of the world's experts on game theory and evolution: John Maynard Smith, professor of biology at the University of Sussex, who submitted TIT FOR TWO TATS. Two people independently submitted REVISED DOWNING. All told 62 entries were received, and generally speaking they were of a considerably higher degree of sophistication than those in the first tournament. The shortest was again TIT FOR TAT; the long-

TAKE THE MATHEMATICS LIBRARY -A $98.85 VALUE- FOR ONLY $2.95





When you join THE LIBRARY OF SCIENCE

You simply agree to buy 3 more books—at handsome discounts—in the next 12 months.

This intriguing, authoritative and thought-provoking 3-volume set with a publishers' price of $98.85 can be yours for $2.95 when you join The Library of Science book club. Here's what you will receive:

CATASTROPHE THEORY FOR SCIENTISTS AND ENGINEERS

By Robert Gilmore.

A massive, invaluable reference for improving the analysis and prediction of change on every conceivable level. Against a background of simple differential calculus, the author presents Elementary Catastrophe Theory from a point of view which is tailored to the needs and backgrounds of physical scientists and engineers. Applications covered include phase transitions in classical and quantum mechanical systems, climatology, optical caustics and diffraction patterns. "An excellent text."—*Nature.*

666 pages. Fully illustrated.
Publisher's Price $49.95.

MATHEMATICAL DISCOVERY: On Understanding, Learning, and Teaching Problem Solving

By George Polya.

A newly combined edition of Polya's two-volume classic. Using a hands-on approach, this step-by-step guide will show you how to solve problems requiring a fair degree of judgment, originality and creativity in algebra, geometry and combinatorics. This heuristic method teaches you how to recognize when a problem is suitable for mathematical treatment, and how to plan a strategy for tackling it.

"This classic text on problem-solving...is still ripe with strategy, insight, heuristics, and hints for teachers."
—*The American Mathematical Monthly.*

432 pages. Softcover.
Publisher's Price $20.95.

THE MATHEMATICAL EXPERIENCE

By Davis and Hersh.

In 60 vivid voyages, the authors, both prize-winning mathematicians, explore the abstract, scientific and aesthetic aspects of mathematics, making understandable such forbidding areas as group theory, the Riemann hypothesis, non-Cantorian set theory and non-standard analysis. "A true gem, one of the masterpieces of our age."—*The American Mathematical Monthly.*

440 pages. Extensively illustrated with charts, diagrams and photos.
Publisher's Price $27.95.

4 GOOD REASONS TO JOIN

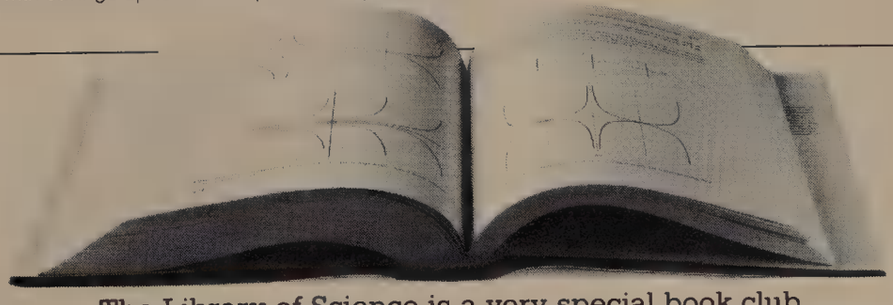
1. The Finest Books. Of the hundreds of science books published each year, only the very finest are selected and offered. Moreover, our books are always of equal quality to publishers' editions, *never* economy editions.

2. Big Savings. In addition to getting The Mathematics Library for only $2.95, you keep saving substantially—up to 30% and occasionally even more. (For example, your total savings as a trial member—including this introductory offer—can easily be over 50%. That's like getting every other book free!)

3. Bonus Books. Also, you will immediately become eligible to participate in our Bonus Book Plan, with savings up to 70% off publishers' prices.

4. Convenient Service. At 3-4 week intervals (16 times per year) you will receive the Book Club News, describing the Main Selection and Alternate Selections, together with a dated reply card. If you want the Main Selection, do nothing and it will be sent to you automatically. If you prefer another selection, or no book at all, simply indicate your choice on the card, and return it by the date specified. You will have at least 10 days to decide. If, because of late mail delivery of the News, you should receive a book you do not want, we guarantee return postage.

If reply card has been removed, please write: The Library of Science—Dept. 2-AW3, Riverside, N.J. 08075 to obtain membership information and application.



The Library of Science is a very special book club.
A club dedicated to keeping members of the scientific community (and knowledgeable amateurs) aware of the important advances and the speculative thinking in their own and related fields.

Enter The New ROBOTICS AGE

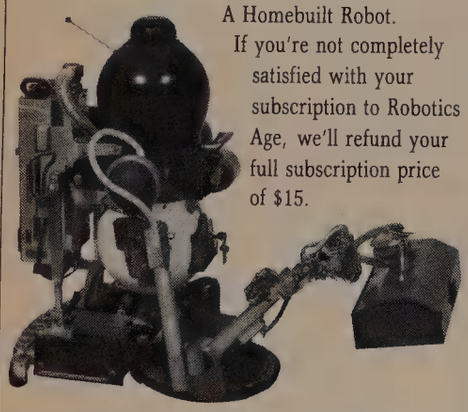
An age where formerly impossible tasks become everyday reality. The mechanization of many laborious physical and intellectual tasks of industrial, commercial and domestic life is coming about through the technology of intelligent machines. We see innovative applications of microprocessors, simulations, sensor electronics, real time control, effector design, machine design, aerospace robots and autonomous systems.

This is the world of Robotics Age, the Journal of Intelligent Machines. Make this world yours. In Robotics Age you'll find topics from the realities of industrial automation to the prospect of personal robotics. Integrating it all is modern software engineering. As illustrations, we show how you can perform practical design explorations with inexpensive personal computers.

Recently published titles include: Avatar, A Homebuilt Robot; The Physics of One-Legged Mobile Robots; Constructing an Intelligent Mobile Platform; An Inexpensive Hand; Natural Language Understanding; Applying Robot Vision To the Real World; Telecommunications Robots. In every issue you'll find features like Patent Probe, New Products, Letters and Design Forums.

Don't miss out: Subscribe today. With your paid subscription accompanied by the coupon below, we'll send you a free reprint of the article Avatar: A Homebuilt Robot.

If you're not completely satisfied with your subscription to Robotics Age, we'll refund your full subscription price of $15.



meet **AVATAR:** A Homebuilt Robot

Sign me up for a 6 issue trial subscription . . .
[ ] Send an issue and bill me $15.00 for 6 in all.
[ ] Enclosed is $15.00 (6 issues plus Avatar reprint)

Name

Address

Town

State Zip

Charge My [ ] Master Card [ ] Visa

Number Expires

Robotics Age Magazine
P.O. Box 358 – Peterborough NH – 03458
(603) 924-7136 **SA**

est was a program from New Zealand, consisting of 152 lines of FORTRAN. Once again RANDOM was added to the field, and with a flourish and a final carriage return the horses were off! Several hours of computer time later the results came in.

The outcome was nothing short of stunning: TIT FOR TAT, the simplest program submitted, won again. What is more, the two programs submitted that had won the subjunctive replays of the first tournament now turned up way down on the list: TIT FOR TWO TATS came in 24th, and REVISED DOWNING ended up buried in the bottom half of the field. This may seem horribly nonintuitive, but remember that a program's success depends entirely on the environment in which it is swimming. There is no single "best strategy" for all environments, so that winning in one tournament is no guarantee of success in another. TIT FOR TAT has the advantage of being able to "get along well" with a great variety of strategies, whereas other programs are more limited in their ability to evoke cooperation.

Axelrod puts it this way: "What seems to have happened is an interesting interaction between people who drew one lesson and people who drew another lesson from the first round. Lesson One was 'Be nice and forgiving.' Lesson Two was more exploitative: 'If others are going to be nice and forgiving, it pays to try to take advantage of them.' The people who drew Lesson One suffered in the second round from those who drew Lesson Two."

Hence the majority of participants in the second tournament really had not grasped the central lesson of the first tournament: the importance of being willing to initiate and reciprocate cooperation. Axelrod feels so strongly about this that he is reluctant to call two strategies playing against each other "opponents"; in his book he always uses neutral terms such as "strategies," "players" and so on. He even does not like saying they are playing *against* each other, preferring "with." I have tried to follow his usage in this article, with occasional departures. One striking fact about the second tournament is the success of "nice" rules: of the top 15 finishers only one (which came in eighth) was not nice. Amusingly, a sort of mirror image held: of the bottom 15 finishers only one was nice.

Several non-nice strategies featured rather tricky probes of the opponent (sorry!), sounding it out to see how much it "minded" being defected against. Although this kind of probing by a program might fool occasional opponents, more often than not it backfired, causing severe breakdowns of trust. Altogether it turned out to be very costly to try to use defections to "flush out" the other player's weak spots. It proved to be more profitable to have a policy of cooperation as often as possible together with a willingness to retaliate swiftly against any attempted undercutting. Note, however, that strategies featuring massive retaliation were less successful than TIT FOR TAT, with its gentler policy of restrained retaliation. Here forgiveness is the key, since it helps to restore the proverbial "atmosphere of mutual cooperation" (to use the phrase of diplomacy) after a skirmish.

The overall lesson of the first tournament was in essence "Be nice and forgiving." Apparently, however, many people just could not get themselves to believe it; they were convinced that with cleverer trickery and scheming they could win the day. It took the second tournament to prove them wrong. And from the second tournament a third key strategic concept emerged: that of *provocability*—the idea that one should "get mad" quickly at defectors and retaliate. Thus a more general lesson is "Be nice, provocable and forgiving."

Strategies that do well in a wide variety of environments are called by Axelrod "robust," and it seems the strategies with "good personality traits"—that is, nice, provocable and forgiving strategies—are sure to be robust. TIT FOR TAT is by no means the only possible strategy with these traits, but it is the canonical example of such a strategy and is astonishingly robust.

Perhaps the vividest demonstrations of TIT FOR TAT's robustness were provided by various subjunctive replays of the second tournament. Of all the types of replay that Axelrod tried (nearly all of which were won by TIT FOR TAT) undoubtedly the most significant and ingenious was the "ecological tournament." Such a tournament consists not only of a single subjunctive replay but also of an entire cascade of hypothetical replays, each one's environment determined by the results of the preceding replay. In particular, if you take a program's score in a tournament as a measure of its "fitness," and if you interpret "fitness" to mean "number of progeny in the next generation," and finally if you let "next generation" mean "next tournament," then what you get is that each tournament's results determine the environment of the next tournament—and successful programs become more numerous in the next tournament. This type of iterated tournament is called ecological because it simulates ecological adaptation (the shifting of a fixed set of species' populations according to their mutually defined and dynamically developing environment) as contrasted with the mutation-oriented aspects of evolution, where *new* species can come into existence.

As one carries an ecological tournament through generation after genera-

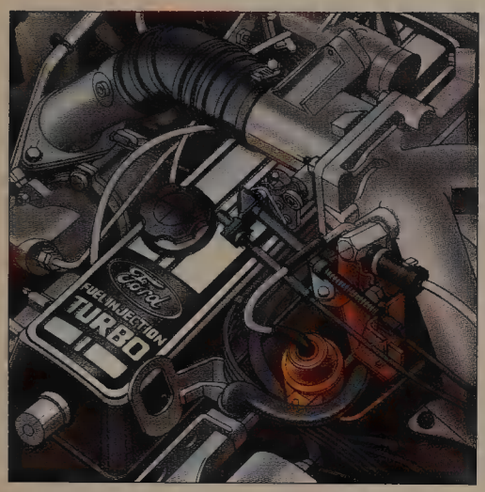
THUNDERBIRD TURBO COUPE.
A WORLD CLASS TOURING CAR
FROM FORD.

THUNDERBIRD TURBO COUPE

Turbo Coupe is a fuel-injected, turbocharged, five-speed Thunderbird with sophisticated running gear, performance compound tires, and seats that are more than a place to sit. A driver's car.

SOPHISTICATED POWER, JUDICIOUSLY APPLIED.

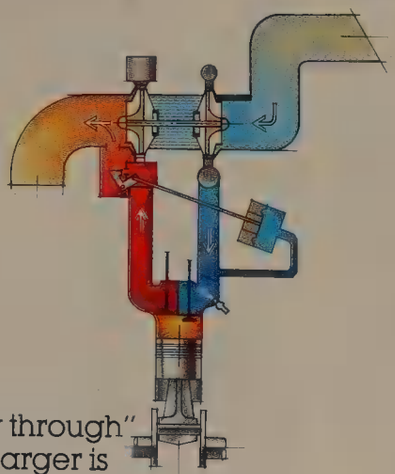
The engine.



An efficient, high-performance cross-flow head, overhead cam four-cylinder displacing 2.3 liters.

The engine employs five main bearings, forged aluminum pistons, high-temperature alloy valve materials, oil cooler and tuned intake manifold.

The turbocharger.



A "blow through" turbocharger is positioned upstream from the injectors and throttle plate.

Unlike other turbocharging arrangements, this "blow through" system pressurizes on demand. Which in turn gives the driver crisp, immediate throttle response. At full boost, around 4,600 engine RPM, the engine achieves 145 horsepower based on SAE Standard J-1349. More power than most of today's V-8s. A triumph of technology over brute force.

The electronic fuel injection system with EEC-IV.

The fuel injection system, fed by an electric fuel pump, is port type—with an injector ahead of each intake valve for precise metering.

The system is run by a computer called EEC-IV. A fourth-generation, state-of-the-art microprocessor-based engine control system capable of thousands of operations per second.

At one extreme, it will immediately set up air/fuel mixture and timing for a quick cold-start. At the other extreme, it will help the engine push to its safe limits. In between, it senses by the millisecond what you are asking the car to do, and balances the engine's air/fuel mixture and ignition timing for optimum power, response and efficiency.

The 5-speed transmission.

The free-revving powerplant is mated to a 5-speed gearbox. The ratios are carefully calibrated so there is a proper gear for any turn, any straightaway.

Accomplished drivers will appreciate the pedal design. The brake and accelerator pedals are positioned to facilitate true heel-and-toe downshifting.

BALANCE AND POISE ON THE ROAD.

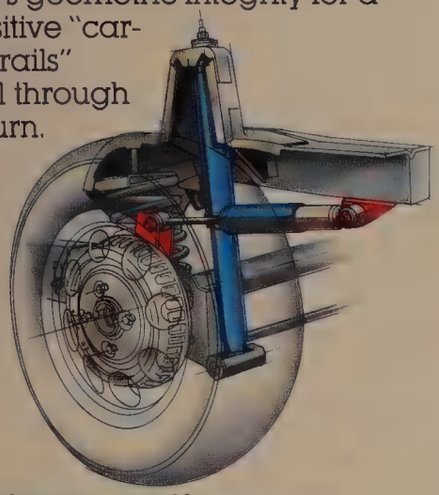
Aerodynamics/air flow management.

Turbo Coupe has one of the lowest drag coefficients of any four-passenger high-performance car in America. But more than that, Thunderbird uses the wind to help hold itself firmly on the road. It reduces lift on the hood for directional stability. Reduces lift on the rear for cornering agility. The smooth lines even lower wind noise inside the cabin. This is the difference between mere aerodynamics and what we call "air flow management."

Gas-filled struts and shocks for ride and handling.

The front struts and rear shocks on Turbo Coupe contain a low pressure charge of nitrogen. This suppresses foaming of the hydraulic fluid due to rapid bumps or exuberant driving. This in turn means the ride doesn't have to be overly stiff in normal driving to be stiff enough when you, or the road, demand.

Plus two additional rear shocks for pure handling.
Turbo Coupe has a second set of rear shocks mounted in a rather unique and interesting position...longitudinally between the rear axle and the body. These help maintain the car's geometric integrity for a positive "car-on-rails" feel through a turn.



Performance tires.
Turbo Coupe mounts Goodyear Eagle P205/70HR 14 performance tires on lightweight cast aluminum wheels. The tires feature an aggressive tread pattern, performance compound rubber, and low profile design for a wide footprint. Result: excellent traction and cornering power on wet or dry pavement.

MAN. MACHINE. UNITY.

Seats you sit in. Not on.
Turbo Coupe's front seats reward the demanding driver. (1) The under thigh support adjusts. (2) The side thigh support adjusts. (3) The seat back angle adjusts. (4) The lumbar support adjusts with (5), an infinitely adjustable pneumatic actuating pump. In addition, the (6) firm side supports help hold you in place. And unlike some other high-performance seats, you don't have to be an accomplished acrobat to get in and out. Considering that a Turbo Coupe is capable of .81G in a turn, seats that support you laterally as well as up and down are essential.

Command centers.
There are three. The driving center puts steering wheel, light controls and shift lever right where your hands want them to be. The environment center clusters radio and climate control a comfortable reach away. The console center nestles electric controls for windows, seats and mirrors in a pod next to your right thigh. The car feels as if it flows from your fingertips.

IN SUM.

An engine so efficient it's rated at 145 horsepower on one hand and 33 estimated highway, (22) EPA estimated miles per gallon on the other.* Superb handling with no surprises.

And a feel that blends man with machine in a thoroughly harmonious way. A driver's car in every respect.

ONE FINAL BIT OF INFORMATION.

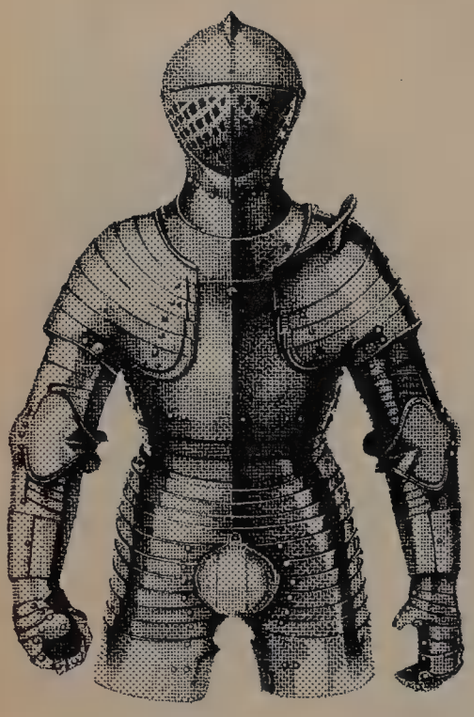
Fewer than 10,000 Turbo Coupes are scheduled to be built this first model year. Now you know what you should do. Quickly.
Call 1-800-772-2100 for more information.

*For comparison. Your mileage may differ depending on speed, distance and weather. Actual highway mileage lower.

Get it together—Buckle up.

HAVE YOU DRIVEN A FORD...LATELY?

Ford



Treasures from the Tower of London

A unique collection of armour housed in the Tower of London now making its first ever journey to Toronto. Over five hundred years of history are detailed in these exquisitely crafted examples of arms and armoury, showing at the newly renovated and expanded Royal Ontario Museum.

The collection of one hundred and twenty works includes suits of armour belonging to Emperor Maximilian I of Germany, and Henry VIII and Charles II of England, as well as ceremonial and presentation pieces, hunting equipment, edged and missile weapons and a variety of firearms.

The exhibition runs from March 19 to June 19. Hours are Monday to Sunday 9:00 a.m. to 11:00 p.m. Admission is $5.50 for adults, $4.00 for seniors and students, and $3.00 for children, 12 and under.

For more information, call TOLL FREE: **1-800-828-8585,** in New York State, **1-800-462-8404.**

ONTARIO yours to discover!

tion the environment gradually changes. At the beginning poor programs and good programs alike are equally represented. As time passes, the poorer ones begin to drop out as the good ones flourish. The rank order of the good ones, however, may now change, because their "goodness" is no longer being measured against the same field of competitors.

Thus success breeds ever more success, but only provided that the success derives from interaction with other similarly successful programs. If, in contrast, some program's success is due mostly to its ability to milk "dumber" programs for all they are worth, then as those programs are gradually squeezed out, the exploiter's base of support will be eroded and the program will suffer a similar fate.

A concrete example of ecological extinction is provided by HARRINGTON, the only non-nice program among the top 15 finishers in the second tournament. In the first 200 or so generations of the ecological tournament, as TIT FOR TAT and other successful nice programs were gradually increasing their percentage of the population, HARRINGTON too was increasing its percentage. This was a direct result of HARRINGTON's exploitative strategy. By the 200th generation or so, however, things began to take a noticeable turn. Weaker programs were beginning to become extinct, which meant there were progressively fewer dupes for HARRINGTON to profit from. Soon the trend became apparent: HARRINGTON could not keep up with its nice rivals. By the 1,000th generation HARRINGTON was as extinct as the dodoes it had exploited.

Axelrod summarizes: "Doing well with rules that do not score well themselves is eventually a self-defeating process. Not being nice may look promising at first, but in the long run it can destroy the very environment it needs for its own success."

Needless to say, TIT FOR TAT fared spectacularly well in the ecological tournament, increasing its lead ever more. After 1,000 generations not only was TIT FOR TAT ahead but also its rate of growth was greater than that of any other program. This is an almost unbelievable success story, all the more so because of the absurd simplicity of the "hero." One amusing aspect of it is that TIT FOR TAT did not defeat a single one of its rivals in their encounters. This is not a quirk; it is in the nature of TIT FOR TAT. TIT FOR TAT cannot defeat anyone; the best it can achieve is a tie, and often it loses (although not by much).

Axelrod makes this point very clear: "TIT FOR TAT won the tournament, not by beating the other player but by eliciting behavior from the other player that allowed both to do well. TIT FOR TAT was so consistent at eliciting mutually rewarding outcomes that it attained a higher overall score than any other strategy in the tournament.

"So in a non-zero-sum world you do not have to do better than the other player to do well for yourself. This is especially true when you are interacting with many different players. Letting each of them do the same as or a little better than you is fine, as long as you tend to do well yourself. There is no point in being envious of the success of the other player, since in an iterated Prisoner's Dilemma of long duration the other's success is virtually a prerequisite of your doing well for yourself."

Axelrod gives examples from everyday life in which this principle holds. Here is one: "A firm that buys from a supplier can expect that a successful relationship will earn profit for the supplier as well as the buyer. There is no point in being envious of the supplier's profit. Any attempt to reduce it through an uncooperative practice, such as by not paying your bills on time, will only encourage the supplier to take retaliatory action. Retaliatory action could take many forms, often without being explicitly labeled as punishment. It could be less prompt deliveries, lower quality control, less forthcoming attitudes on volume discounts or less timely news of anticipated market conditions. The retaliation could make the envy quite expensive. Instead of worrying about the relative profits of the seller, the buyer should worry about whether another buying strategy would be better."

Like a business partner who never cheats anyone, TIT FOR TAT never beats anyone, yet both do very well for themselves.

One idea that is amazingly counterintuitive at first in the Prisoner's Dilemma is that if the other player is unresponsive, the best possible strategy for you to follow is ALL D. It might seem some form of random strategy might do better, but that is completely wrong. If I have laid out all my moves in advance, then playing TIT FOR TAT will not do you any good, nor will flipping a coin. You should simply defect on every move. It does not matter what pattern I have chosen. Only if I can be influenced by your play will it do you any good to cooperate.

Fortunately in an environment where there are programs that cooperate (and whose cooperation is based on reciprocity) being unresponsive is a very poor strategy, which in turn means ALL D is a very poor strategy. The single unresponsive competitor in the second tournament was RANDOM, and it finished next to last. The last-place finisher's strategy was responsive, but its behavior

was so inscrutable that it *looked* unresponsive. And in a more recent computer tournament conducted by Marek Lugowski and me in the Computer Science Department at Indiana University three ALL D's came in at the very bottom (out of 53), with a couple of RANDOM's giving them a tough fight for the honor.

One way to explain TIT FOR TAT's success is simply to say it elicits cooperation by means of friendly persuasion. Axelrod spells this out as follows: "Part of [TIT FOR TAT's] success might be that other rules anticipate its presence and are designed to do well with it. Doing well with TIT FOR TAT requires cooperating with it, and this in turn helps TIT FOR TAT. Even rules that were designed to see what they could get away with quickly apologize to TIT FOR TAT. Any rule that tries to take advantage of TIT FOR TAT will simply hurt itself. TIT FOR TAT benefits from its own nonexploitability because three conditions are satisfied:

"1. The possibility of encountering TIT FOR TAT is salient;

"2. Once encountered, TIT FOR TAT is easy to recognize; and

"3. Once recognized, TIT FOR TAT's nonexploitability is easy to appreciate."

This brings out a fourth "personality trait" (in addition to niceness, provocability and forgiveness) that may play an important role in success: recognizability, or straightforwardness. Axelrod chooses to call this trait clarity, and he argues for it with clarity: "Too much complexity can appear to be total chaos. If you are using a strategy that appears random, then you also appear unresponsive to the other player. If you are unresponsive, then the other player has no incentive to cooperate with you. So being so complex as to be incomprehensible is very dangerous." How rich these comments are in their implications about social and political behavior!

In a letter to Axelrod, Rapoport cautioned against overstating the advantages of TIT FOR TAT; in particular he believes the strategy is on occasion too harshly retaliatory. It can also be persuasively argued that on other occasions TIT FOR TAT is too lenient. Certainly there is no evidence that TIT FOR TAT is the ultimate strategy. Indeed, as has been emphasized repeatedly, the very concept of "best" is incoherent, since all depends on the environment. In the tournament at Indiana University several strategies like TIT FOR TAT did better than pure TIT FOR TAT. They all shared, however, the three critical "character traits" whose desirability had been clearly revealed by Axelrod's earlier analysis of the important properties of TIT FOR TAT. They were simply a little better at detecting nonresponsiveness, and when they were convinced

YOUR EXCITING NEW HOBBY!

- Enjoy fantastic savings by assembling your own organ or piano.
- It's easy. No technical knowledge required.
- Just follow our clear, pictured instructions.
- Choose from many models from portables to consoles.
- Ask about our interest-free installment plan.

WERSI
ORGAN & PIANO KITS

WERSI Dept. M41 P.O. Box 5318
Lancaster, PA 17601

☐ Free Info. Pack: ☐ Organ ☐ Piano
☐ Catalog & Demo Record *(enclose $3)*

Name _____
Address _____
City _____ State _____ Zip _____
Phone ( ) _____

Rep inquiries invited

LYNCHBURG
HARDWARE & GENERAL STORE®
47 Main St., Lynchburg, TN 37352

JACK DANIEL SQUARE GLASS SET

Mr. Jack Daniel was the originator of the square bottle for his whiskey and always wanted to have a matching square glass. Well, here it is! This hefty square glass (each weighs 14 ounces) is the perfect companion to a bottle of Mr. Jack's finest. The inside is rounded to make drinking a pleasure and the original design is fired on for good looks and durability. My **$15.00** price for a set of 4 glasses (8 oz. capacity) includes postage.

Send check, money order or use American Express, Diners Club, Visa or MasterCard, including all numbers and signature. (Add 6¾% sales tax for TN delivery.) For a free catalog, write to Eddie Swing at the above address. Telephone: 615-759-7184.

C65 SPOTTING SCOPE
A Revolution in Outdoor Optics from Celestron®

At Your CELESTRON DEALER

The C65 is one of Celestron's growing number of precision optical products that brings you *adventure.*
The Celestron C65 introduces modern mirror-lens technology to spotting scope users. The C65's folded mirror-lens optics eliminates the need for bulky porro prisms that absorb light and come out of adjustment in conventional spotting scopes. Celestron's C65 spotting scope offers you powerful true-color images of the world around you, all from a lightweight instrument less than a foot long.
Please write or call (toll-free) for information. Visit a Celestron dealer and look through the revolutionary C65, it will change your world. With a Celestron you won't have to wait for adventure ... you can look for it.

PRECISION OPTICS from CELESTRON®
Call Toll Free 1-800-421-1526 for Information
(M-F 8AM-4:30PM Pacific Time within the continental U.S.)
Celestron International P.O. Box 3578, SA, 2835 Columbia St. Torrance, CA 90503 (213) 328-9560

¡Speak Spanish like a diplomat!

What sort of people need to learn a foreign language as quickly and effectively as possible? *Foreign service personnel,* that's who. Members of America's diplomatic corps are assigned to U.S. embassies abroad, where they must be able to converse fluently in every situation.

Now you can learn to speak Spanish just as these diplomatic personnel do — with the Foreign Service Institute's Programmatic Spanish Course. You'll learn Latin American Spanish recorded by native speakers.

The U.S. Department of State has spent thousands of dollars developing this course. It's by far the most effective way to learn Spanish at your own convenience and at your own pace.

The course consists of a series of cassettes and accompanying textbook. Simply follow the spoken and written instructions, listening and repeating. By the end of the course, you'll be learning and speaking entirely in Spanish!

This course turns your cassette player into a "teaching machine." With its unique "programmatic" learning method, you set your own pace — testing yourself, correcting errors, reinforcing accurate responses.

The FSI's Programmatic Spanish Course comes in two volumes, each shipped in a handsome library binder. You may order one or both:

☐ **Volume I: Basic.** 12 cassettes (17 hr.), manual, and 464-p. text, $125.
☐ **Volume II: Intermediate.** 8 cassettes (11½ hr.), manual, and 614-p. text, $110.

(Conn. and N.Y. residents add sales tax)

TO ORDER BY PHONE, PLEASE CALL TOLL-FREE NUMBER: **1-800-243-1234.**

To order by mail, clip this ad and send with your name and address, and a check or money order — or charge to your credit card (AmEx, VISA, MasterCard, Diners) by enclosing card number, expiration date, and your signature.

The Foreign Service Institute's Spanish course is unconditionally guaranteed. Try it for three weeks. If you're not convinced it's the fastest, easiest, most painless way to learn Spanish, return it and we'll refund every penny you paid. Order today!

81 courses in 26 other languages also available. Write us for free catalog. Our 10th year.

Audio-Forum
Suite 13G
On-the-Green,
Guilford, CT 06437
(203) 453-9794

AUDIO-FORUM®

the other player was unresponsive, they switched over to an ALL D mode.

In his book Axelrod takes pains to spell out the answers to three fundamental questions about the evolution of cooperation in a world of raw egoism. The first concerns *initial viability:* How can cooperation get started in a world of unconditional defection—a "primordial sea" swarming with unresponsive ALL D organisms? The answer (whose proof I omit here) is that an invasion by small clusters of cooperating organisms, even if they form a tiny minority, is enough to give cooperation a toehold. One cooperator alone will die, but small clusters of cooperators can arrive (by means of mutation or migration, say) and propagate even in a hostile environment, provided they are defensive like TIT FOR TAT.

The second fundamental question concerns *robustness:* What type of strategy does well in unpredictable and shifting environments? We have already seen that the answer to this question is any strategy possessing the four fundamental "personality traits" of niceness, provocability, forgiveness and clarity. This means such strategies, once established, will tend to flourish, particularly in an ecologically evolving world.

The final question concerns *stability:* Can cooperation protect itself from invasion? Axelrod proved that it can indeed. In fact, there is a gratifying asymmetry to his findings: although a world of "meanies" (ALL D strategies) can be penetrated by cooperators in clusters, a world of cooperators *cannot* be penetrated by meanies, even if they arrive in clusters of any size. Once cooperation has established itself it is permanent. As Axelrod puts it: "The gear wheels of social evolution have a ratchet."

The term "social" does not mean these results necessarily apply only to higher animals that can think. Clearly four-line computer programs do not think, and yet it is in a world of just such "organisms" that cooperation has been shown to evolve. The only "cognitive abilities" needed by TIT FOR TAT are recognition of previous partners and memory of what happened the last time with a particular partner. Even bacteria can do this, by interacting with only one other organism (so that recognition is automatic) and by responding only to the most recent actions of their "partner" (so that memory requirements are minimal). The point is that the entities involved can be on the scale of bacteria, small animals, large animals or nations. There is no need for "reflective rationality"; indeed, TIT FOR TAT could be called "reflexive" (in the sense of being as simple as a knee-jerk reflex) rather than "reflective."

For people who think that moral behavior toward others can emerge only when there is some horrendous and totally external threat (of fire and brimstone, say) or some soothing promise of heavenly reward (such as eternal salvation) the results of this research must give pause for thought. Axelrod captures the entire idea in one sentence: "Mutual cooperation can emerge in a world of egoists without central control, by starting with a cluster of individuals who rely on reciprocity."

There are so many situations in the world today where these ideas seem of extreme relevance—indeed urgency—that it is tempting to draw all kinds of morals. In the later chapters of his book Axelrod offers advice about how to promote cooperation in human affairs, and at the end the political scientist in him cautiously ventures some broad conclusions about global issues, which are a fitting way for me to conclude:

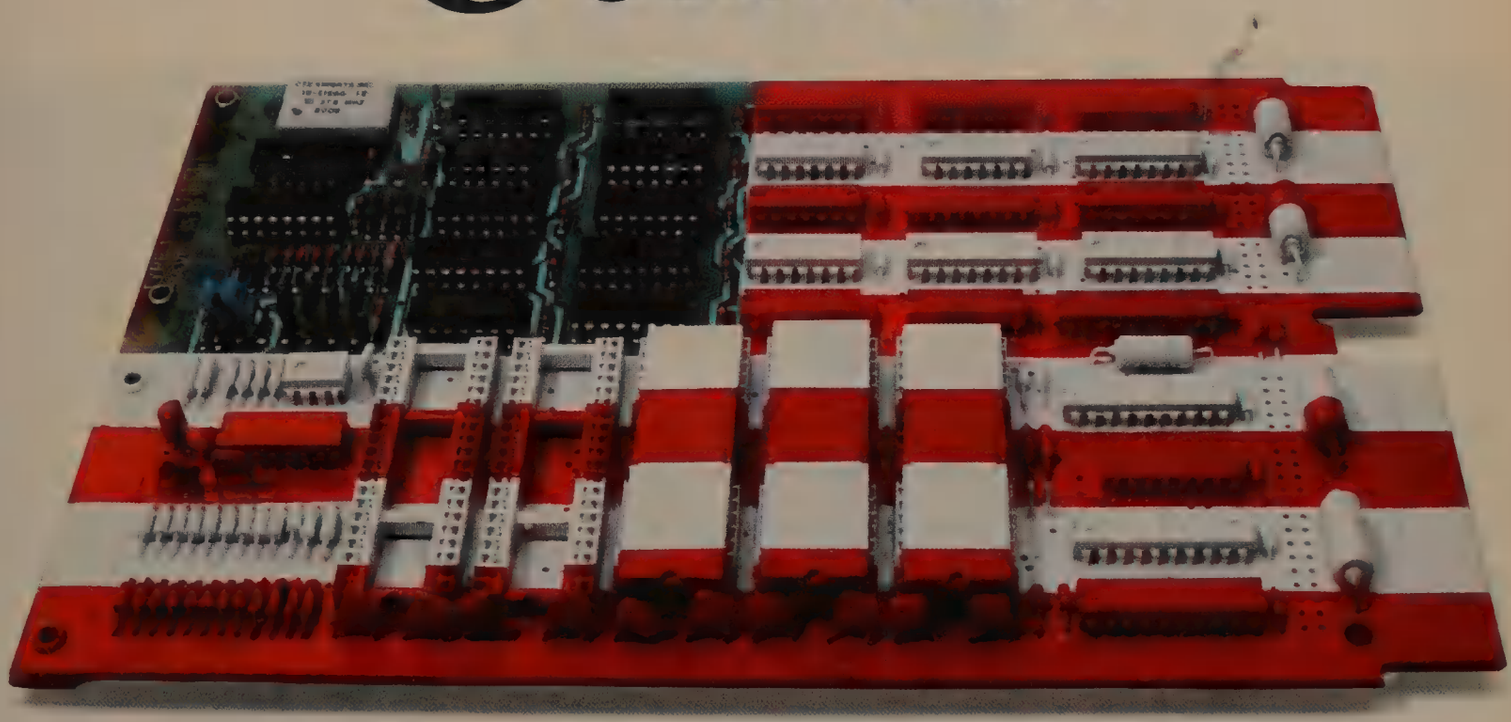
"Today, the most important problems facing humanity are in the arena of international relations where independent, egoistic nations face each other in a state of near anarchy. Many of these problems take the form of an iterated Prisoner's Dilemma. Examples can include arms races, nuclear proliferation, crisis bargaining and military escalation. Of course, a realistic understanding of these problems would have to take into account many factors not incorporated into the simple Prisoner's Dilemma formulation, such as ideology, bureaucratic politics, commitments, coalitions, mediation and leadership. Nevertheless, we can use all the insights we can get.

"Robert Gilpin [in his book *War and Change in World Politics*] points out that from the ancient Greeks to contemporary scholarship all political theory addresses one fundamental question: 'How can the human race, whether for selfish or more cosmopolitan ends, understand and control the seemingly blind forces of history?' In the contemporary world this question has become especially acute because of the development of nuclear weapons.

"The advice given in this book to players of the Prisoner's Dilemma might also serve as good advice to national leaders as well: Don't be envious, don't be the first to defect, reciprocate both cooperation and defection, and don't be too clever. Likewise, the techniques discussed in this book for promoting cooperation in the Prisoner's Dilemma might also be useful in promoting cooperation in international politics.

"The core of the problem is that trial-and-error learning is slow and painful. The conditions may all be favorable for long-run developments, but we may not have the time to wait for blind processes to move us slowly toward mutually rewarding strategies based upon reciprocity. Perhaps if we understood the process better, we could use our foresight to speed up the evolution of cooperation."

Advanced Technology Centers



Strategies for Corporate Growth

Innovation, development, and mature manufacturing at the edge of knowledge need different kinds of support. Where to find that support is basic to corporate strategy.

Peter J. Brennan and
Development Counsellors International, Ltd.

Advertisement

"During the next five years, Michigan will be the home for venture capital."

That's what Dwight Carlson, founder and president of Perceptron, a Michigan machine vision manufacturer, has to say about Michigan's massive new capitalization initiatives.

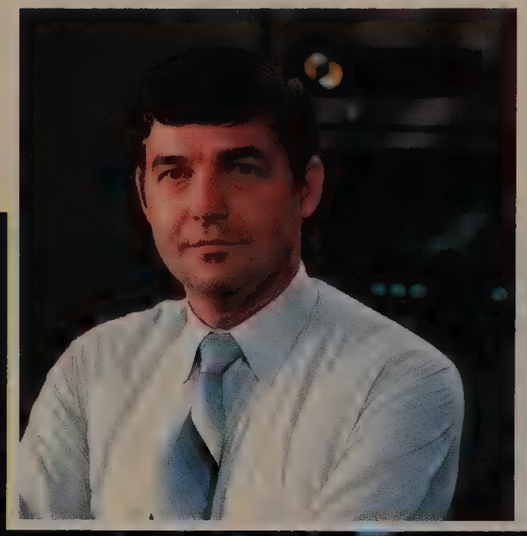
Thanks to the liberalization of retirement fund usage, and the establishment of the Michigan Economic Development Authority and the Michigan Certified Development Corporation, businessmen like Mr. Carlson will find more than five hundred million new dollars in Michigan. Michigan offers new financing for venture capital, land acquisition, machinery, construction and conversion. A state research fund provides grants for Michigan-based entrepreneurs to develop their products.

"Michigan has improved my chance for success and allowed me to build here," says Carlson. "I thought I'd face a lot of red tape and programs with no substance. But in less than three months, I've negotiated a substantial investment from the state retirement fund along with some of the top venture capital firms in the country."

Innovative capital programs like these, along with Michigan's exceptional research and educational facilities and its skilled labor force, move Carlson and colleague James Swartz of Adler & Company, a leading venture capital firm, to say: "We're excited about the positive action being taken in Michigan to enhance the growth of technology firms."

Let the consultants in Michigan's Office of Economic Development help you Say Yes with financing, location and state-financed labor training. To find out more call toll free: 1-800-I SAY YES (in Michigan 1-800-292-9544) or simply write on your business letterhead and mail to:

Albert A. Bogdan, Director
Office of Economic Development
Michigan Department of Commerce
Dept. 10583, P.O. Box 30225
Lansing, Michigan 48909



Dwight Carlson, President, Perceptron

Say Yes to Michigan!

Attend: Bio-Connections '83, July 14 and 15, 1983.

Business Opportunity in Applied and Exploratory Biology
Michigan State University, East Lansing, MI
For information contact: Paul R. Stuhmer, Lansing Regional Chamber of Commerce,
Box 14030, Lansing, MI 48901. (517) 487-6340.

Innovation, development, and mature manufacturing at the edge of knowledge need different kinds of support. Where to find that support is basic to corporate strategy.

It was 1938. David Packard was a young engineer, newly graduated from Stanford University, working for one of the world's largest technology firms. His employer was General Electric Company in Schenectady, New York, a pleasant and long-established industrial city a continent away from the sunny Santa Clara Valley, between the bay and the sea.

The Valley had two main exports—fruit and college graduates. At Stanford, the late Professor Frederick Emmons Terman, a tireless promoter of his students, of Stanford, of the Santa Clara Valley and of the beginnings of electronic technology embodied then in radio, fretted over the loss of his brightest students. The Professor found them jobs, too often somewhere else like Schenectady. When he could, he found financing and research contracts to keep his graduates in California, close to home.

Thus William Hewlett stayed on to do graduate work. His friend, David Packard, married to a Californian, could be persuaded to return on a fellowship. Hewlett invented an audio oscillator and Terman, himself a transplanted Midwesterner, suggested Hewlett and Packard try marketing it. In early 1939, they did, from the garage of Packard's rented house.

From Orchards to Technology

Thus did the Valley begin to change to an economy based on new technology. It was not planned. There were no consultants' studies. No one was persuaded to shut down an uneconomic plant somewhere else and move it to cheaper California. No economic development organization promoted Palo Alto to all parts of the country.

Rather, as is true for most successful advanced technology centers, a local, possibly unique, combination of people, facilities and circumstances combined to yield an unforeseen result.

Stanford had excellent electrical engineering and physics departments, poised at the edge of practical applications for electronics technology. In Fred Terman, the school and the locale had a teacher, an organizer, a man of vision and determination, a superior technologist with a firm grasp of the real world. He was the essential industry-university-community linkage.

David Packard, a Coloradan, was prepared to go to the University of Colorado but ". . . the time I spent in Palo Alto in the summer of 1929 convinced me that I should apply to Stanford."

Like most human adventures, the process begins at a point no one then present can identify. Through a series of circumstances and serendipities, it gathers momentum. The phenomenon acquires shape, becomes large enough to have its own center of gravity. Like a black hole it begins to attract and coalesce new ingredients. A new-technology center appears amid the prune yards, on the swampy land above a salt bed, on the prairie, strung out along a highway . . .

Most important, the world notes that the new technology center is a net gain for both its region and the nation as a whole. Like a seed dropped on the forest floor, it has germinated, grown to maturity, gains its subsistence from roots much deeper than the surrounding shallow plants to which it gives shelter and sustenance. In time, it drops its own seeds, which grow to form a grove.

Cultivating Technology Business

We long ago learned how to farm trees so that we no longer depend on accidents of nature for timber. We can analyze after the fact what prompted a center of advanced technology to germinate and grow in the wild. We have been less successful in transplanting those conditions so that we can start and grow new technology centers wherever we want.

Technology is the key to increasing productivity of both capital and labor. A producer made most efficient by computer-aided design (CAD), state-of-the-art instrumentation and control, electronic data processing for inventory and materials management and modern telecommunications is not as vulnerable as his less technology-intensive, less efficient competitor.

Advanced technology is thus vital to the competitiveness of older industries, which will continue to exist and expand. Whether they reindustrialize in the United States or migrate entirely to other parts of the world largely depends on how well they adapt new technology to their own ends. Survival also hangs on how well the high-technology industries themselves fill the needs of their non-high-technology so-called smokestack industry customers.

State governments recognized long ago that advanced technology is the key to future prosperity. The economic vitality of Silicon Valley and the Minneapolis-St. Paul area are well known. The virtual rebirth of the Massachusetts economy based on seven high-technology industries that often started out in the abandoned factories of long-gone smokestack industries has not gone unnoticed. Neither has the symbolism. Recent economic history, however, has greatly concentrated the attention of both state governments and businesses.

Traditional regional, state or area development programs have aimed at painting an attractive picture for any kind of industry. Development officers largely were marketers rather than developers. Their sales message was and often still is financial incentives. The long-term objective was an increase in local and state tax bases and simultaneous increases in employment.

The strategy works well enough in a rising economy when addressed to traditional non-high-technology industries, whose requirements are based on the products they make, the raw materials they need and the markets they serve. However, if a rising tide raises all boats, it is equally true that a falling one exposes the rocks and impales a few of the larger vessels, which will never float again. More than any other in the past twenty years, the 1979-83 recession exposed the rocks in the American industrial economy.

High Tech Is Different

The strategy does not work as well when addressed to high technology industries. Measured against the factors that traditional industries must consider, high technology industries are virtually site-independent. They can set up just about anywhere. And they do, all too often.

Area development officers must ask: who are the Hewlett-Packards of today and what do they need?

Companies that help corporations choose sites for new facilities have started to find those answers. Robert M. Ady, executive vice president of The Fantus Company, a Chicago-based site location firm, says: "When we deal with a traditional industry, the company's short list of preferences will all be in the same region—say Texas, Oklahoma and Louisiana. But when we deal with high technology companies,

the preferences will be scattered across the country—California, Texas and North Carolina."

States, regions and localities have started to recognize this fundamental difference between high technology and traditional industries. Further, they recognize that there are differences within the high-technology industries themselves. The most technology-intensive of all industries, guided missiles and spacecraft (Standard Industrial Classification [SIC] 376) has some of the aspects of a traditional industry. At its manufacturing level, this industry deals with big items that need plenty of space. Among the largest single manufacturing facilities in the world are aircraft plants in Texas, Kansas, and coastal Washington.

Hi Tech— a Definition

According to the Bureau of Labor Statistics, a high technology manufacturing firm is one in which engineers and scientists comprise more than five percent of the total work force. That, however, is too loose a definition. It includes transportation equipment in general, for example, of which aerospace is a part, and chemicals, of which pharmaceuticals are a part.

Advanced technology industries are better defined as those that require high levels of continuing innovation and whose markets can change overnight. These firms typically have 10 percent or more scientists and engineers. Firms working at the leading edge of technology have 15 percent or more engineers and scientists on staff.

Only six industries qualify. These and their SIC codes are: Pharmaceuticals (283), Computers (357), Semiconductors (367), Communications (367), Aircraft (372), and Instruments (381 through 384).

Instruments includes medical, controlling and scientific instruments.

One must also include the service sectors that are so increasingly important in the U.S. economy. More than most manufacturing sectors, such service industries as finance, banking, communications, software development, insurance, medical services and data-processing are fast-growing users of advanced-technology products and the driving force behind the rapid commercialization of new ones. Commerce is more and more a market for industry—but commerce basically services industry.

A high-technology company typically goes through three distinct phases. Fantus's Robert Ady defines these as the theory-driven, product-driven and market-driven stages. A successful innovative and productive company will exist on all three levels. However, each phase requires a different set of circumstances at its beginning. As each phase becomes a continuing level, it requires still another set of conditions to prosper and grow.

The Three Drives

First is the initial scientific discovery and the follow-on work that converts the discovery to a commercial product. A lone and brilliant inventor, a Hewlett or an Edison, may well make the initial discovery working in his basement far from any known technology center. He may even carry the discovery to the point of commercial viability. But as the frontiers of knowledge have expanded, wresting new secrets from science and moving them up the scale to commercial practicality has become an expensive effort.

The second product-driven stage entails development and first-level manufacturing. Once a scientific idea has been proved and its practical applications divined and defined, a company will exploit it. The company may be a new one formed out of the basic research group that developed the idea, or it may be a large established one seeing opportunity in new technology.

The classical view in the United States is that small, entrepreneurial firms bring most new technologies to market. And in fact, most new jobs produced in the U.S. economy come from small companies exploiting new technologies. However, large companies with large R&D budgets and extensive facilities spend far more on R&D than do small new firms. So do universities, which, says the National Science Foundation (NSF), perform half of all basic research in the U.S.

During the product-driven stage, a company depends on its source of basic science and technology, which may be the company itself. More likely, though, the source will be a nearby university, research institute or large technology-oriented corporation from which the company's founders came. As the company expands and increases its R&D expenditures, it will become less directly dependent on external technology. Ideally, it will itself become a technology source in a growing technology center.

Outgrowing Home Base

As the growing firm's manufacturing capacity expands, its need for space and other services grows. Manufacturing remains closely linked to research and development sources. Consequently, if manufacturing and R&D cannot be at the same place, any new site must be within easy reach of the R&D center. The manufacturing site must have locally available nearly the same level of technology as does the home site. To attract and hold engineers and scientists, the new location must offer similar amenities.

Few people, not Hewlett, not Packard, not Samuel N. Irwin, deliberately move to a location specifically to found a company. "Companies start where people are," says Mr. Irwin, founder and president of Irwin International in Ann Arbor, Michigan. The founders may be there because of a university, an employer, research foundation, government installation or were born there. Having started the firm, local factors keep the founders in the area, at least through initial success and early expansion.

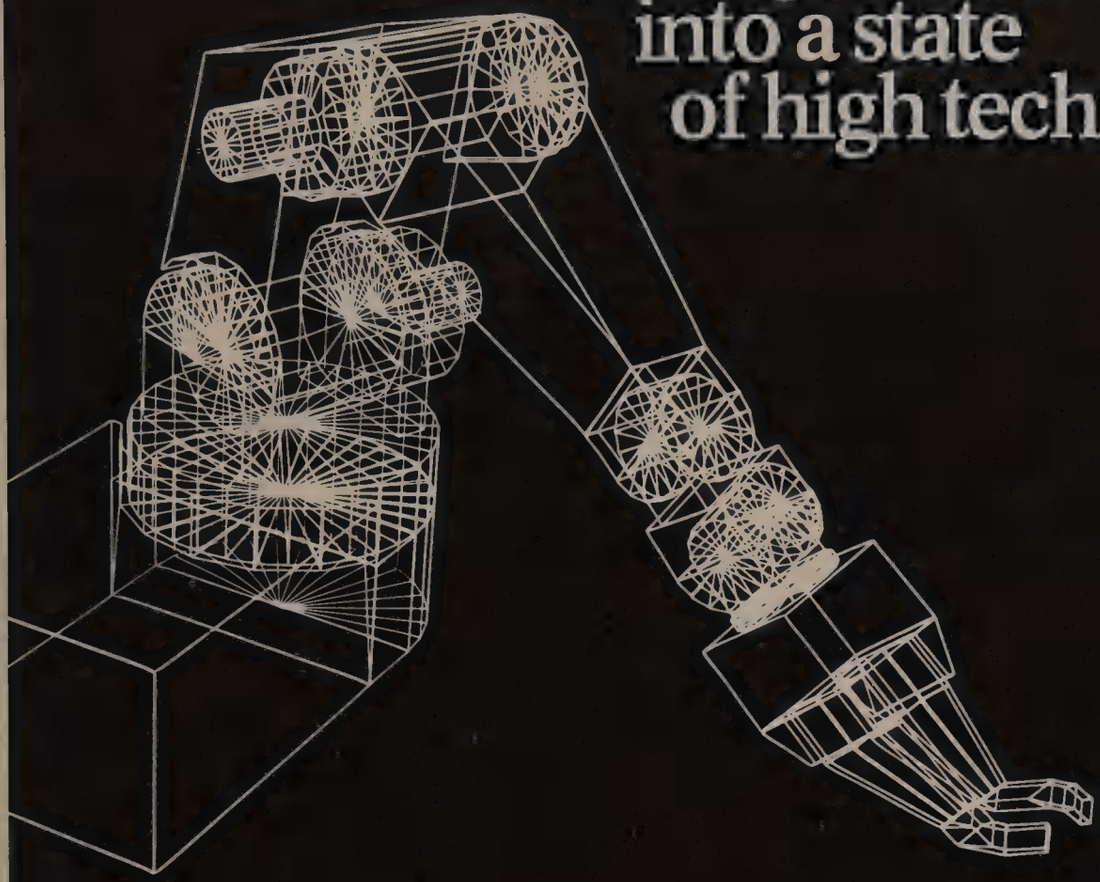
Local factors may enable a newborn company to grow to a certain point in its product-driven phase. However, this stage is a critical time in a company's life. Further expansion may produce new needs that the local economy cannot fill. In its early years, any fast-growing company is locked in to its location by its need for cash to finance growth. It must generate most of that cash internally. Venture capitalists may fund a company only through startup. Equity markets want to see a track record. Conventional bankers prefer more established, less risky ventures.

When the Company Should Move

When a company outgrows the local supply of brain power, cannot attract the people it needs, encounters too much red tape in continuing financing, it moves to an area where the community understands and is prepared to serve its needs.

Or a company may outgrow its manufacturing space. Its first satellite is likely to be nearby in an area that offers most or all of the benefits of the headquarters location. Since the com-

Work yourself into a state of high technology.



Utah.

Illustration courtesy GE CAE International and Evans & Sutherland Computer Corp.

Advanced State of Technology

Utah has assembled all of the components for the advanced technology industry to thrive. Things like first-class educational and research facilities. Capable engineers and scientists. An abundance of skilled workers and technical assistance support. A business-boosting economy. And a quality of life that is second to none.

Decades of proven performance

Many of this century's most creative innovators have pioneered their work in Utah. Breakthroughs in bio-engineering like the implantation of artificial human organs. Highly advanced computer simulators to train astronauts for outer space. Sophisticated micro-electronics systems to dramatically reduce the time it takes to design large, integrated circuits. All of these efforts and more are products of Utah's unique cooperative research community—a marriage of education, science, government, industry and the working people of Utah.

Technology isn't all Utah has advanced

Utah has a wide variety of cultural, recreational and educational activities inviting participation. The state is rich in four season, outdoor attractions, plus Utah blends the rustic and scenic wonders of nature with a cultured and cosmopolitan environment.

Experience Utah!

Economic and Industrial Development Division
200 South Main Street, Suite 620, Dept. SA
Salt Lake City, Utah 84101
Phone: 1-800-453-8824

Tell me more about the state of high technology in Utah. Please send me: ☐ your new advanced technology brochure ☐ a comparative study report on the costs of doing business in Utah vs. 12 other major markets ☐ a list of companies in Utah I can contact.

I am particularly interested in: _____
(type of industry)

Name _____ Title _____

Company Name _____ No. of Employees _____

Product/Service _____

Address _____

City _____

State _____ Zip _____

pany can now choose its new site, management will look more closely at factors the founders could not control.

Taxes Are Important

There is a misconception that tax structure is important only to large-scale, large-employment smokestack traditional industries. To justify this view, people point to California and Massachusetts, the top advanced-technology states but far from the bottom of the list on overall tax burden.

Not so. "Taxes are very important to small, expanding high-technology companies," says Dr. Robert Premus, Dayton-born staff economist of the Subcommittee on Monetary and Fiscal Policy of the Joint Economic Committee. Premus is the author of a recent report entitled "Location of High-Technology Firms and Regional Economic Development," based on a survey of 691 companies.

"Small companies in particular rank the tax burden high," said Dr. Premus in an interview. "Cash flow is critical. Local taxes can take cash from a company when it most needs it." At the same time, they are locked into their location. By the time they can afford to move, the tax burden has become secondary. They may even move to a more heavily taxed jurisdiction.

The founder of one high-technology company explained that paradox. William C. Norris, chairman of Control Data Corporation in Minnesota, said that the State of Nebraska's willingness to increase both the tax base and rate was a powerful factor in the company's decision to locate a new facility there. "Nebraska demonstrated that it meant to provide the tangible and intangible facilities and services that (we sought)," said Norris. "Without such a change in philosophy," the Nebraska native continued, "we likely would have gone elsewhere."

The Market-Driven Facility

The third stage identified by Robert Ady is the mature manufacturing or market-driven level. The company and its products have now grown to a point where costs are paramount. The company has outgrown its older facilities. Processes and products are mature. Transportation, local financial incentives and tax structure, availability of low cost and trainable labor all assume greater importance in the site selection decision.

The market-driven firm's products are high technology. But manufacturing is essentially an assembly process. The technology and science is in the maintenance of the manufacturing systems and in quality control of raw materials and finished items. Engineers and scientists are a small percent of total employment.

This third stage plant will have different needs than the earlier stage one. This new facility is still not tied much to natural resources, nor overly dependent on market proximity. Such site-independent plants can serve markets in North America as easily from Taiwan or Malaysia as they can from Tennessee. Beyond minor input by local managers, all technology is generated elsewhere. New technology arrives daily, by telephone and computer network or encased in the latest model of manufacturing systems equipment—but it comes from somewhere else.

Such are the plants that site development managers seek. They are clean, reasonably kind to the environment. Their demands on the local infrastructure are easily met if there is adequate electrical power and water. They provide jobs for local people without placing great demands on the local services and amenities. They do not bring in large numbers of people who are accustomed to more than the community provides.

However, since these plants are essentially assembly operations, they commonly do not have the same ties to their location as does the company's technology base. To quote Dr. Robert Premus's Congressional report: "The survey . . . indicate(s) that high-technology companies are 'footloose' . . . access to raw materials . . . markets and transportation are not major locational determinants. Nor are . . . water . . . energy . . . and climate important determinants . . . high-technology companies are drawn more to highly specialized resources such as labor skills and education and to factors that make it easier to attract and maintain a skilled labor force, most notably State and local taxes . . ."

The survey also indicated that most high-technology companies prefer an urban to a rural environment. The centripetal effect of an urban-centered location brings in more people and companies and encourages technology transfer. Indeed, if a university in the traditional sense is a community of scholars, then a high-technology center may itself be a university—a community of scholarly companies.

Dynamics of High-Tech Industry

Most states historically understood little about the dynamics of high-technology industry. To attract such companies, they did little analysis and less development. In effect, the state rounded up what already existed and packaged it attractively for itinerant industry. The package did little for resident businesses and less still for nascent ones.

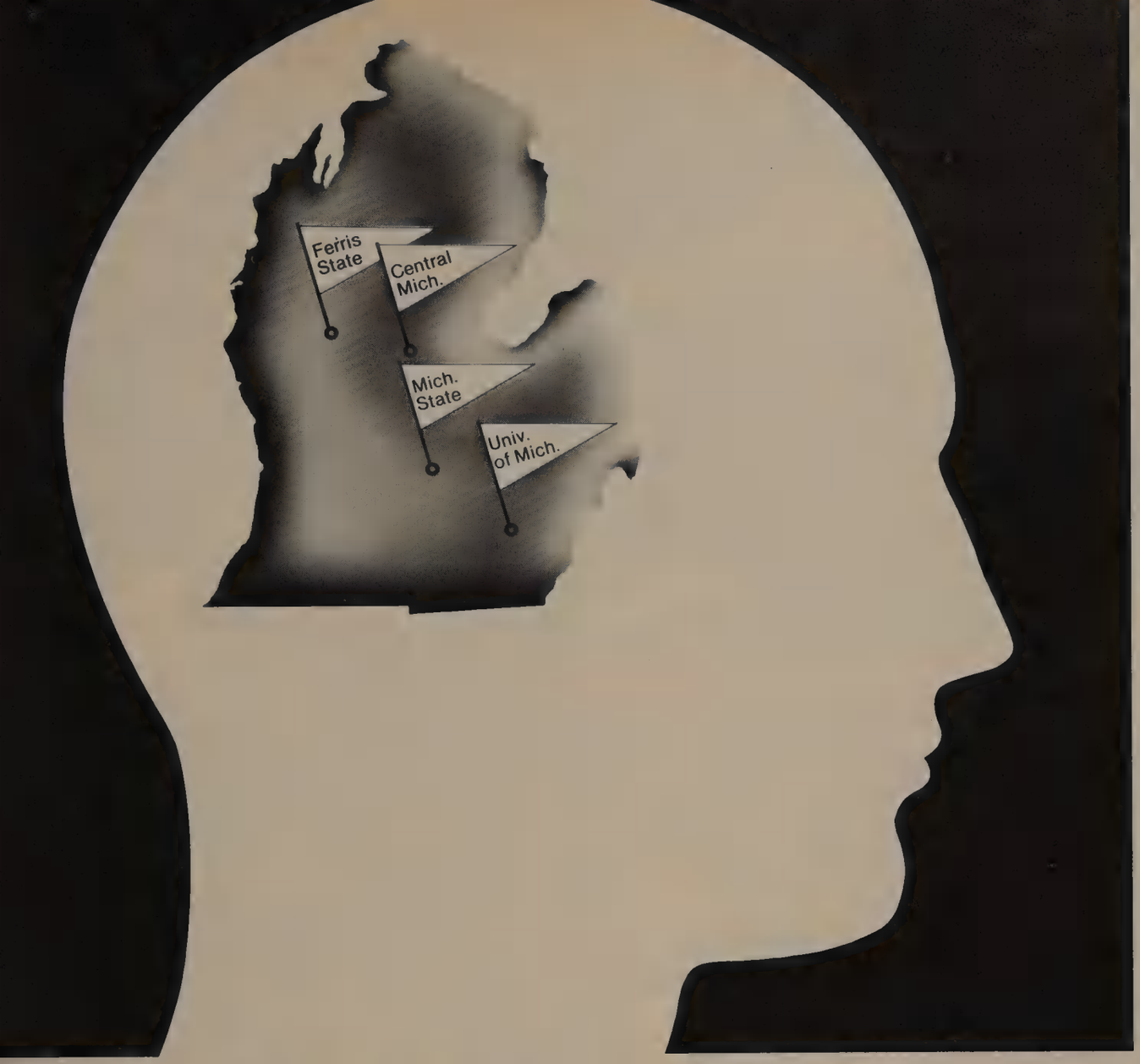
Miles Friedman, Executive Director of the National Association of State Development Agencies (NASDA), notes the changes in the attitudes of state development people. "It is still true that many state agencies see it as their jobs to move plants from other states to theirs. But a common business complaint is that the state agencies ignore the businesses they already have. Now there is big emphasis on in-state development and the incubation of new businesses."

Some states and localities years ago tried to develop conditions that would permit advanced technology centers to flourish. Not until 1981 did the National Governors Association (NGA) establish a Task Force on Technological Innovation. Local generation of high-technology industry is clearly an idea whose time has come.

Partially as a result of the NGA's initiative, many states from Maine to Hawaii established high-technology study groups. Eventually, some states will take all possible steps to establish and strengthen the structural elements needed to breed high-technology industry. Some few may decide that their future economies will depend on maintaining the major base of their present economies. The states' new awareness of high technology can help direct technological resources to appropriate economic sectors, whatever their place on the technology spectrum.

The Governors' Survey

In late 1982, the NGA Task Force surveyed every state governor on organization of state efforts, economic



MICHIGAN

If you want to get technical, Michigan is the place to be.

For high-powered, high-tech industries, Michigan is precisely the place to be. Businesses on the move will find $500,000 of venture capital available here. What's more, our universities are among the finest in the country. And to a technically oriented business looking for a place to grow, that can mean prompt access to nationally prominent consultants. And proximity to advanced research and testing facilities.

Vehicle research and crash testing. Agriculture. Medicine. Robotics. Genetic engineering. They're all areas where ongoing research and advancements have won Michigan universities world renown. But that's just a part of the high-tech catalyst. Michigan is also home for numerous technical organizations: the Center for Research in Integrated Manufacturing, the Michigan Technology Council, the Industrial Technology Institute, the Molecular Biology Institute and the Robot Institute of America.

Consumers Power knows that Michigan is where science and education get down to business. And we know the results of this coalition. If your high-tech business needs a place to succeed, let us put you in touch with those in Michigan who make things happen – technically speaking.

You may discover that surrounding yourself with Michigan's technical and educational resources is the quickest way to put your business in a class by itself. If you're technically oriented, call Richard R. Surbrook at (517) 788-0344. Or mail the coupon.

Consumers Power

Say yes to Michigan!

For more information, mail this coupon to:
Richard R. Surbrook
Director of Economic Development
Consumers Power Company
212 West Michigan Avenue, Jackson, MI 49201

Name

Company

Address

City/State/Zip SA0583

WANTED.

PARTNERS TO HELP PENNSYLVANIA FORGE THE FACTORY OF THE FUTURE.

You have a dream, a vision of the coming technology.

Its seeds were planted by the likes of Asimov, Clarke and H.G. Wells. Its growth nurtured by every teacher, class, professor and experiment that ever challenged you. And inspired you.

In Pennsylvania we, too, have a dream. The creation of an environment in which our universities, industries and state government join forces in a partnership to help scientists, engineers and entrepreneurs like you forge your dreams into the technologies of the future.

Much of that environment already exists. As you read this, more than 150,000 Pennsylvania engineers and scientists are developing all facets of advanced technology. Robotics. Microelectronics. CAD/CAM. Biotechnology.

There are now more major research universities in Pennsylvania than any other Northeastern state. And those schools are supported by corporations whose R&D contributions are among the highest in the nation.

Last year we awarded almost nine million dollars in low interest loans to advanced technology companies.

We also created the catalyst that will make our dream a reality.

The Ben Franklin Partnership.

It combines the talents and resources of our universities, small and large corporations, and state government to further stimulate advanced technology research and development. Millions of dollars in grants will provide funds for everything from basic product research to specialized work force training and sophisticated market planning.

As thorough as this partnership is, it lacks one essential element.

You. Your talent. Your vision.

Send the coupon to Walter Plosila, Deputy Secretary for Technology. Or call him at (717) 787-3003. You'll learn more about the quality of commitment, as well as the quality of life you will find here.

In Pennsylvania, we share your dream.

Come, be partner to ours.

Mail To: Department of Commerce, 433 Forum Building, Box 2001, Harrisburg, PA 17120

Name _____ Title _____

Company _____

Address _____

Telephone _____

ADVANCED TECHNOLOGY HAS A FRIEND IN PENNSYLVANIA

© 1983, Pennsylvania Department of Commerce Dick Thornburgh, Governor James O. Pickard, Secretary of Commerce

incentives, local university-industry cooperation and worker training programs. The survey report in July 1983 will have two sections. One organized by state will describe state policies and programs in technological innovation. The other, organized by program category, will make it easy to compare state approaches to particular aspects of economic development and technology, by various methods.

Among the early findings: at least 11 states have appointed task forces or boards as overall policy-determining bodies on technology. These are California, Illinois, Iowa, Kansas, Michigan, Maryland, New Jersey, North Carolina, Ohio, Pennsylvania and Virginia. Most set these up in 1982, though North Carolina was way ahead of its time, having established its Board of Science and Technology, chaired by the Governor, in 1963.

Twenty-two states have advisory groups outside the state government. Their scope and linkage to the government varies greatly. Georgia, for example, recently set up the Advanced Technology Center at Georgia Institute of Technology. The Governor and legislature created the Center in 1980, but it is not a government body.

State money and private-sector matching funds finance the Maine Development Foundation, which the state legislature enabled in 1977. A governor's advisory committee proposed the cooperative business and government approach in 1975. In Minnesota, twenty-eight prominent people established last year the Minnesota Wellspring. The governor is honorary chairman. New Hampshire, which has developed a high-technology service center based on publishing and computer software around the town of Peterborough, has the Center for New Hampshire's Future. This is a private organization established in 1979.

The not-for-profit Indiana Corporation for Science and Technology was authorized by the General Assembly in 1982. The Governor appoints a 24-member board representing the private, public and educational sectors. California has the Commission on Industrial Innovation; Maryland its Governor's Advisory Council; Michigan the proposed High Technology Corporation; New York the Science and Technology Foundation; Pennsylvania the Governor's Council on Science and Technology; and South Carolina a proposed Industrial Research Board. The City of Chicago and The State of Illinois, together with Chicago universities, have created an "Illinois Technology Partnership."

The Early Starter

Few programs were as early or as ambitious as North Carolina's. As George Herbert, president of North Carolina's Research Triangle Institute describes it, the state in the mid-fifties was near the bottom in per capita income and too dependent on "old-line segments of its economy: agriculture, textiles, tobacco manufacturing, furniture and brick and tile." At the same time, the state ranked among the top ten in numbers of colleges and universities. But for their graduates there were few local opportunities.

North Carolina looked enviously at the centers in Massachusetts and California and noted that the desired industry existed in areas of strong research concentration. Further, most important research centers lived in the shadow of major graduate-level universities. The state had at least three such schools close together: Duke at Durham; the University of North Carolina at Chapel Hill; and North Carolina State University at Raleigh.

The state in 1956 decided to establish a research park in the triangle formed by the universities, thus the Research Triangle. The concept was not an instant success, nor did the founders expect it to be. As Mr. Herbert says, "They knew they were working for returns that would not be truly significant for 15, 20 or 25 years." And he cautions newcomers to the high technology stakes, "This is a reality, too often ignored by many of the groups that visit us today looking for a quick fix for 1983's economic woes."

No one really knows the optimum size and mix of the critical mass that turns a high-technology manufacturing center into a truly innovative one. Few doubt that it takes many years to reach that self-igniting concentrated mix of research centers, universities and theory-oriented industry.

The planners first established the Research Triangle Foundation to acquire land and develop the 5,700-acre park. They also set up the Research Triangle Institute in close association with the universities. None of these is a state agency.

Despite an early capture of a major industrial laboratory in 1959, others were slow to follow. By 1965, there were only nine laboratories with a total employment of 1,000. But then things began to pick up. IBM bought 400 acres. The National Institutes of Health established its National Institute of Environmental Health Sciences. Another major industrial laboratory moved in and the project was off and running.

The most recent addition is the state-sponsored Microelectronics Center of North Carolina, established in 1980. The legislature appropriated the first one million dollars for startup, and the state is providing the first 24.4 million dollars for ongoing support.

North Carolina's experience has many lessons for others. But it is not yet clear that the state has managed to produce an industrial rival to Silicon Valley or Route 128. The location is high on everyone's short list when looking at possible new locations. The universities and RTI itself are well-reputed basic research centers. But the bubbling ferment of entrepreneurial activity that is the true mark of a self-generating advanced-technology center is not yet evident.

A Downtown Research Park

Another research center started on a different premise but at about the same time is the University City Science Center in Philadelphia, Pa. There, the University of Pennsylvania and Drexel University were situated in deteriorating neighborhoods. The universities, business and community conceived The Science Center as a way to improve the city's technology base and reverse urban blight. Under the auspices of some 28 member institutions that own it, the Center began in 1964 in a renovated building. It is now an urban research center covering some 19 acres with over seventy science-based organizations in its nine buildings. Some of these organizations are new to the city while, more important, the Center encourages others to remain there.

The University City Science Center is so far the nation's only downtown research park, though Detroit and Wayne State University in Michigan are starting a similar concept in the Detroit-based Metropolitan Center for High Technology.

The University City Science Center has a relatively small Research Institutes Division. The Research Triangle Institute is a much larger, free-standing organization similar to such older institutions as Battelle Memorial Institute at Columbus, Ohio; Stanford Research Institute in Menlo Park, Calif.; Midwest Research Institute in Kansas

City, Kansas; IITRI in Chicago; Southwest Research Institute in San Antonio, Texas, and many more.

For smaller companies and even larger ones faced with scientific questions that they have neither time nor staff to handle themselves, the research institutes are an important resource. As do most companies, the research institutes often specialize in technologies that reflect their geographical location and the interests of their major clients.

Subdivisions for Technology

Clean high-technology industry usually does not need large and heavy installations. One can develop industrial parks for such industry just as one develops residential sub-divisions for people. A well laid out industrial or research park with room for expansion can be an important factor in site selection, particularly for a smaller, high-technology company making its first move out of its basement.

Successful developers present a balanced facility. If the park itself does not contain a research institute and university campuses, these will be within easy reach. The park will also have conference facilities, hotels, shopping centers and restaurants. A large development may be a completely planned community.

The Huron Center near Ann Arbor, Mich., is such a real estate development. A joint venture between Mitsubishi and Morgan Stanley, the Center is a 393-acre multi-use development eight miles from the University of Michigan and 25 miles from Detroit. When completed, it will contain residences as well as research labs, hotels as well as light industrial plants.

Utah, recently in the news for the artificial heart work at the University of Utah Medical Center, has a growing advanced-technology center in what one publicist refers to as "Bionic Valley," near Salt Lake City.

Montgomery County, Maryland, capitalizes on the nearby concentration of Government medical research facilities embodied in the National Institutes of Health, Bethesda Naval Hospital and many medically-oriented companies in the area. The county has established the 232-acre Shady Grove Medical Park, which has reserved 145 acres for medical science-related businesses and institutions.

Also in Montgomery County is the

IN HIGH-TECH NEW JERSEY, WE NOT ONLY WORK HARDER ...WE WORK SMARTER!

New Jersey has more scientists, engineers, and technicians per capita than any other state in the country. It's also the number one state in private R&D funding. And almost one in every ten research dollars spent nationally is spent in New Jersey's 700 industrial and academic laboratories.

New Jersey is where you'll find the leaders in aerospace research and telecommunications... leaders in the development of faster and smaller computer chips and the development of plastics lighter than aluminum and stronger than steel... leaders in virtually every high-tech field.

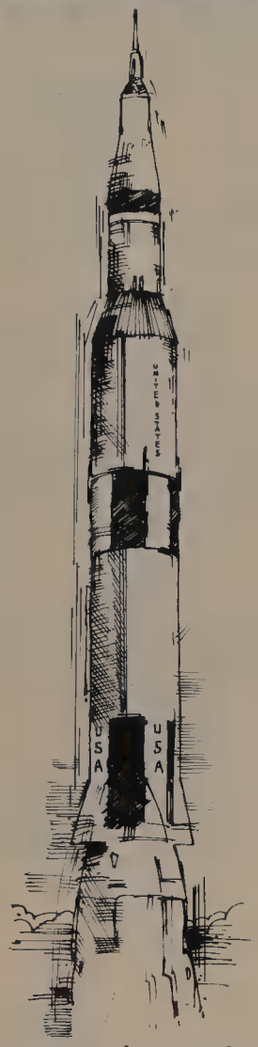
SA13-83

FREE NEW JERSEY FACT BOOK

If you're a young research-oriented firm or established innovative company, there are many reasons why you'd be better off in New Jersey. Learn more about New Jersey and its high-tech environment. For a free copy of the "New Jersey Fact Book" and other material, write: Steve Kukan, Area Development, PSE&G, 80 Park Plaza, Newark, N.J. 07101.

PSE&G Public Service Electric and Gas Company

Technology is in a state.



Tennessee is closer to the moon than it looks: much of the technology launched into space was researched in Tennessee. But space is just one side of Tennessee's technology story. Electronics and computer science are major industries. The University of Tennessee is becoming a nationally recognized energy research center, and we're making important breakthroughs in medical technology. When it comes to high technology, Tennessee is getting down to brass tacks and giving you the moon, and more. For more information, write or call Mike DuBois, Tennessee Department of Economic and Community Development, Andrew Jackson Bldg., Box 999, Nashville, TN 37219. 1-800-251-8594.

Tennessee

We get down to Brass Tacks.

planned city of Columbia with both residences and work places for people in high technology. Columbia mirrors to some extent the planned city of Reston across the river in Fairfax County, Virginia, another site for much high-technology industry.

The Planned City

An ambitious development that has learned much from Columbia, Reston and the Research Triangle is The Woodlands near Houston, Texas. This project of the Mitchell Energy Company is a wholly integrated 25,000-acre residential, research and light manufacturing community. The developers intend that everyone who works in the Woodlands can afford to live there. Several thousand people already do. In that, the community differs little from the traditional concept of the small, self-contained town, which it is.

The research forest has four elements. There will be a 400-acre campus for the University of Houston. The Texas Medical Center Inc., has 150 acres for a research campus. One hundred acres belong to the Houston Area Research Center (HARC), a research institute under the auspices of Texas A&M, Rice University and the University of Houston. Some 1,300 acres are set aside for high-technology businesses and their suppliers.

While developers take pristine land and turn it into parks, in many parts of the country suitable sites lie fallow. These are deactivated military bases. Though these sites are frequently off the beaten path, companies can turn their existing infrastructure to commercial use at low cost. Many were air bases. In Chippewa County, Michigan, for example, market-driven firms can set up on 41 acres of lighted concrete.

Other states and regions hope to use what they already have to generate new technology industry. Illinois, for example, has the Argonne National Laboratory, Fermilab, the Illinois Institute of Technology Research Institute (IITRI), the many fine universities around Chicago and the University of Illinois at Champaign-Urbana. A technology corridor extends westward from Chicago toward Fermilab. The area has an enormous base in existing medical and electronics technology. It has some of the more active and venturesome venture capitalists. But it has lacked the university-industry-community links that have been so important to theory-based entrepreneurial activity elsewhere, which explains recent Task Force development of significant education-based local cooperation.

The GOCO Connection

Many localities with government high-technology installations now see that these facilities are part of the community's technology base. They employ thousands of scientists and engineers as well as equipment and instrumentation that only the wealthiest of industries and few universities can afford. These are the GOCO (Government-Owned, Contractor-Operated) facilities. Much of their technology is in the public domain.

A government high-technology facility linked to local universities creates a high-technology center all its own. With proper encouragement, availability of capital, extended relationships with local businesses, GOCOs become a factor that expanding high technology companies must consider as well, as potent centers for local development.

Tennessee, recognizing the enormous science and technology base at the Union Carbide-operated Oak Ridge National Laboratory and the University of Tennessee a few miles west at Knoxville, has undertaken a major program to develop a technology corridor. Tennessee is not widely thought of as a high-technology area. Yet there are over 2,000 Ph. D. level professionals in the Oak Ridge area alone, including the largest concentration of doctorate-level biologists in the world.

Similar GOCOs and even GOGOs (Government-Owned, Government-Operated) facilities across the country provide ready-made advanced technology centers to which aspiring companies might well attach themselves or from which new ones might spring. Dayton, Ohio, for example, rightly considers Wright Patterson Air Force Base, the Air Force's major technical center, to be a community high technology resource.

The benefits to California of such GOCOs as the Jet Propulsion Laboratory and the Livermore Laboratory are substantial. Florida's technology has gained from the space flight operations at Cape Canaveral and the industries that have sprung up to serve them.

NASA's contribution to the technology base in Houston is incalculable. Joseph P. Loftus, Director of Technical Planning at NASA, Houston, recalls the moon-shot days when thousands of eager engineers and scientists flocked to Houston to be at the leading edge of

something exciting. "I see 15,000 engineers and scientists who once worked here and stayed. They have effected an enormous transfer of material and management technology from the very leading edge where NASA works into the everyday business of the region."

New Mexico has university-operated Los Alamos National Laboratory and AT&T-run Sandia National Laboratory. These and other GOCO facilities of the Department of Energy work in many economically significant areas such as geothermal research, synfuel, solar and new and renewable energy, as well as nuclear fission and fusion.

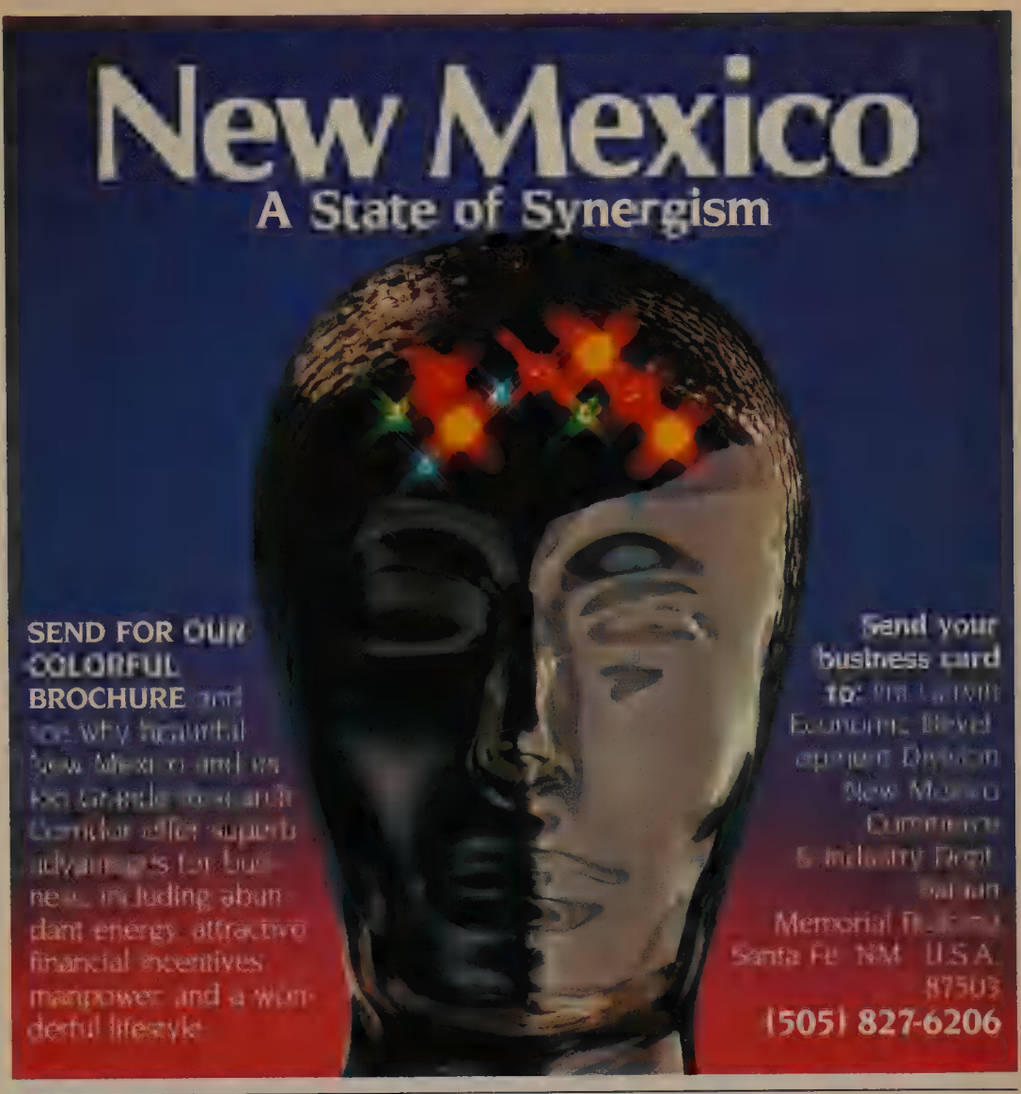
The Federal government's primary laboratory for western coal and lignite research is at Grand Forks, North Dakota. It is now owned by its former contract operator, the 100-year-old University of North Dakota, and renamed the University of North Dakota Energy Research Center. It will continue to handle government projects but will also conduct programs for other sponsors. The Center could be the nucleus of a technology center.

University-Industry Linkages

Successful technology transfer from research to commerce generally results from good linkages between the universities and industry. Such linkages do not always exist. In maintaining something of an ivory tower attitude toward the real world, some universities have forbidden their faculties to engage in commerce. But as manufacturing technology approaches levels of science once found only in the academic research laboratory, and as only industry and government can afford the equipment needed to probe the outer reaches of knowledge, more universities participate in the marketplace.

At the first meeting of the NGA task force, Dr. George A. Keyworth, Science and Technology Adviser to the President, remarked about the problem: "The resistance of the flow of technology from our research laboratories to the marketplace is abysmal," he said.

"There was a decoupling of universities and the business community," says Robert Premus. "The universities sought Federal money for their programs, which were for Federal purposes. That reduced the flow of technology into U.S. industry. But the government cut back support to universities. We see the universities getting anxious and talking to the



Advertisers in this special SCIENTIFIC AMERICAN advertising supplement will be pleased to provide additional information about their services:

Commonwealth of Pennsylvania,
Department of Commerce

Consumers Power Company

Glenrothes Development Corporation

Huron Center

Industrial Development Board
for Northern Ireland

New Jersey Department of Commerce
& Economic Development

Public Service Electric & Gas Company

State of Michigan,
Office of Economic Development

State of New Mexico,
Economic Development Division

State of Tennessee,
Department of Economic & Community Development

State of Utah

State of Vermont

Interested readers may contact advertisers directly or may write: Manager—Industrial Sections, SCIENTIFIC AMERICAN, 415 Madison Avenue, N.Y., NY 10017.

Huron Center

Set Your Site On The Land of Hi-Tech

Huron Center is a prestigious 393 acre development located in the heart of the U.S. high-technology applied industry marketplace. Strategically located on Interstate 94 in Southeast Michigan, the center gives you easy access to the respected research centers of Ann Arbor and the diverse manufacturing facilities of Detroit.

Abundant resources include a large skilled labor force and exceptional support in research and development and high-technology from area academic centers. Local and state governments are eager to lend support.

Let us take the lead in helping you with site selection, design and construction. Contact:

Gary Menzel
Huron Center
769 Medinah
Rochester, MI 48063
(313) 375-2330

Huron Center is a joint venture between Mitsubishi Estate New York, Inc. and Morstan Development Company, the real estate affiliate of the investment banking firm, Morgan Stanley of New York.

ADVANCED TECHNOLOGY

CENTERS—USA

"Michigan—Technology That Works," September 1982.

"Oregon: Environment for Technology," May, 1982.

"Advanced Technology in Illinois," September 1981.

"Advanced Technology Center Santa Clara Valley, California," March 1981.

"Dayton: Formula for High Technology," January 1981.

"Technology of Maryland," December 1980.

"Minnesota: Technology Wellspring," October 1980.

"Advanced Technology Center Fairfax County," September 1980.

For information about obtaining reprints, write:

Manager-Industrial Sections
SCIENTIFIC AMERICAN
415 Madison Avenue
New York, N.Y. 10017

corporate community again."

A respected university has value to an expanding firm, because it supplies a talented labor pool, is an educational and technology resource for the company's staff, and because it attracts a continuing flow of talent. Indeed, basic to technology transfer is the flow from other university-centered areas of people who like birds carry seeds of technology sown elsewhere.

Many people recognize the problem of attracting development support from academia and are doing something about it. Solutions range from organized industry-university linkage organizations through business incubation programs at universities to loosening the rules so that faculty may more easily talk to business.

The University of Wisconsin at Madison has had the University-Industry Research Program (UIR) since 1966. UIR helps develop relationships between UW faculty and industry and helps industry and faculty identify programs, facilities and people with mutual research interests.

The Oregon Graduate Center, a non-university degree granting organization, performs a similar function. The Center has been an important factor in the state's success in attracting spinoff advanced technology companies.

In Pennsylvania, the aptly named Ben Franklin Partnership program concentrates on patents. College and university patent policies cover new developments through the advanced technology centers. Policies concern licensing in the state, royalties and other fees to support advanced technology centers.

Incubating New Businesses

At Carnegie Mellon University in Pittsburgh, Prof. Dwight M. Baumann has been trying a somewhat different approach—the incubator. He set up the Center for Entrepreneurial Development at Carnegie Mellon in 1971. Since a common feature of successful advanced technology centers is a growing cluster of entrepreneurial companies passing on through several generations, Prof. Baumann wanted "to see if we could create a cluster."

Prof. Baumann's Pittsburgh Center has been responsible for 16 first-generation high-technology companies and 9 second-generation ones. One sold for $6 million. Prof. Baumann acknowledges the importance of venture capital, but not necessarily where the entrepreneur is: "Venture capital follows innovation—not precedes it," he points out.

Where The Money Is

Most states recognize that venture capital and follow-on financing are important. Some have innovative programs to channel funds to new companies. Connecticut has a unique concept in the Connecticut Product Development Corporation (CPDC). This publicly-funded, state-chartered corporation's mandate is to invest in innovative new products. CPDC borrows from the state to share the costs of development with the company. CPDC earns royalties and repays the state treasury. The funding is not a loan, but neither is it equity. On a successful product, CPDC can recoup five times the development funding. If the product fails, CPDC loses its money, so far about 3 percent of its total investment.

Michigan has begun to move into areas where private capital has been absent. The state has venture capitalists, notably Doan Resources of Midland, but they are not sufficient to the demand. So the state legislature passed new laws that allow state retirement funds to make equity investments in Michigan businesses. Over $350 million is thus available to high technology companies.

Under new legislation proposed by Governor Ted Schwinden, Montana will invest 25 percent of certain tax proceeds in new and expanding Montana firms, about 13 million dollars the first year. Some 20 percent of the state's own investment portfolio will be in new or expanding Montana firms, about 140 million dollars per year. The state will also create a private-sector venture-capital Montana Development Credit Corporation.

Maryland takes a somewhat more conventional approach. The Development Credit Corporation of Maryland (DCCM) does not take an equity position; rather it lends money to fledgling firms that have shown some evidence of managerial ability to operate at a profit.

Different States, Different Styles

As the competition to attract advanced-technology industry intensifies among advanced technology centers,

the states and regions have begun to resemble the firms they seek. Economic development departments are becoming analogous to profit centers. They raise money, mount marketing efforts, devote more attention to such quality controls as cutting out bureaucratic red tape and honestly analyzing themselves to isolate specific benefits and competitive advantages.

The mode and philosophy of economic development varies greatly from state to state, as the NGA survey makes clear. The governor himself, however, is the chief economic development officer. States Fantus's Robert Ady "The Governor is the ultimate spokesman. He makes a profound impression by his personal involvement in the economic development process."

Some states maintain a relatively low-key, low-budget effort. There may be a small office buried in another department or attached to the Office of the Governor. Cities and counties may have only the business-funded local chamber of commerce.

Others rival many countries. A well-funded central organization may have a cabinet-level director and control of a huge budget for everything from advertising to financial aid. Such an organization may well have branch offices in other states and in foreign countries.

Some states that believe they have a special focus maintain relatively small organizations. Delaware, for instance, seeks high-technology industry but is more oriented toward technology-using service industries like banking and finance, a route North Dakota has also taken. Delaware is one of only four states with a court devoted entirely to business law, the Court of Chancery. Other states, says Delaware, can change their systems to match. But nothing can match Delaware's eighty-year history of case law in business affairs.

A few states have publicly-funded municipal and county development organizations that are nearly as prominent or more so than the state organization. A gain for the county or city is of course a gain for the state. But once a county or city organization learns of a prospect, the rivalry and competition can be as fierce as any between nations.

Utilities Eager to Help

Companies need adequate and dependable utilities. Among the factors that influenced the government's decision to install Fermilab's huge particle accelerator in Illinois, for example, was Commonwealth Edison's ability to satisfy enormous instantaneous power demands at a reasonable price.

Dependable power is seldom a problem in the United States. In any case, utilities are not a deciding factor for most high-technology companies. However, in many states the utility companies are a valuable resource in another way. They are literally wired in to their state or region. As investor-owned companies, they are invaluable information sources.

Sometimes the utilities do more state promotion than the states. As John H. Maddocks, General Manager for Area Development at New Jersey's Public Service Electric and Gas Company puts it: "We're part of New Jersey. The vitality of our company is related to the economy of the state."

Michigan's Consumers Power Company gives prospective Michiganders extensive information. Not only does the company help select a site, its training program for Michigan communities shows them how to retain industry and attract new ones.

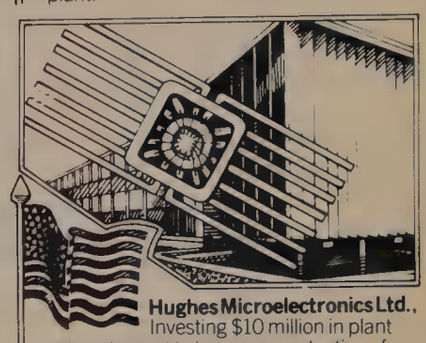
A new company or an expanding one thus has many places to turn in seeking those factors that will best suit its unique needs. Company management, particularly when looking for new sites for market-oriented plants, cannot easily rule out any location. It is a truism that every place will uniquely match someone's needs. And, to quote Florida's promotional motto, "People like to work where they like to live."

Venture Capital— High-Tech Nutrient

Most government financial incentives have little to do with getting the shaky embryonic firm with a great idea out of the basement, garage or university laboratory and into the marketplace. That function is truly the province of venture capital. A few states have made some provision to help guide state funds into venture funding. But it is not yet clear what form these initiatives will eventually take.

They might be repaid out of profits to an extent beyond the initial sum. Such a form, similar to that of the Connecticut Product Development Corporation, is a grant whose reimbursement is contingent on the success of the enterprise. Maximum reimbursement is limited to five times the

General Instrument Microelectronics, New York. Investing $17 million to double the size of their micro-chip fabrication plant.



Hughes Microelectronics Ltd., Investing $10 million in plant and equipment to increase production of microelectronic components and sub-system assemblies.

Your company can be our next Scottish success story.

The four companies above are just a small slice of our impressive track record in Glenrothes. They came and they're flourishing because Glenrothes gives:

- ☐ Britain's biggest development and tax incentives
- ☐ Immediate factory accommodation
- ☐ Highly educated workforce
- ☐ Best industrial climate in Western Europe
- ☐ Easy access to the entire Continent

And that's not all. Glenrothes offers:

- ☐ Excellent housing
- ☐ Five first class universities and two technical colleges
- ☐ Some of the best recreation in Scotland
- ☐ Including 100 golf courses – like St. Andrews and Gleneagles – less than an hour away.

To get the information on how to become our next Scottish success story clip this ad to your letterhead/business card and mail to: **John McCombie, Commercial Director, Glenrothes Development Corporation, 733 3rd Avenue, Suite 2100, NYC 10017** or call collect – (212) 687 1383.

Glenrothes

THE EUROPEAN SUCCESS ROUTE

amount of the original grant.

Venture capital is equity money invested in the early stages of a company's life to help it grow beyond startup. While the initial money will have come from principals and their families, venture capital generally comes from outsiders, who take shares in the company and will become part owners.

Relations between venture capitalists and company founders are complex. They can be difficult depending on how much control the venture capitalist demands and the founder is willing to yield. (For an excellent discussion on this point, see "Raising Venture Capital: An Entrepreneur's Guidebook," published by the New York accounting firm Deloitte Haskin & Sells.)

In the Last Analysis

A management's decision on where to put the new high-technology manufacturing plant can rest on intangibles. Among the factors Bob Premus's survey found was the company founder's place of birth. As says Colorado-born David Packard, "I have always had a high regard for the University of Colorado and it may have something to do with my interest in our location in that part of the country."

Hewlett-Packard's first expansion beyond Palo Alto was to Colorado.

Managements want to increase profits. Mature managers will consciously examine every facet of their company's position within its own evolving industry and the company's own stage of development. Those in the theory-driven phase may also make all the right moves more out of instinct than reasoned strategy—one reason so many company founders find themselves ousted as the company enters the product- and market-driven phases.

But instinctive or planned, any high technology firm must encounter, consider and deal with these many factors governing the right location at the right time in the company's history. No advanced technology business in today's fast-changing markets can stand entirely alone.

A U.S.-based high technology company in the theory- or product-driven stage would not usually locate its operations overseas. There are exceptions, especially in bio-medical technology. An overseas location can allow such a company to bring a product to market sooner under perhaps more relaxed regulatory procedures than those in the U.S.

When a high-technology product reaches the market-driven commodity stage, however, and costs and market access outweigh most other factors, management must look as closely at overseas sites as domestic ones. A company will usually consider off-shore manufacturing if it also has or expects to develop markets abroad. However, foreign production occasionally makes economic sense even when all output goes to the U.S. market.

Many countries are eager to attract high-technology manufacturers. Countries want these industries for the same reasons the states and localities of the U.S. do. They provide jobs, help upgrade local skills and eventually increase the tax base. More important in the view of many countries is the potential for technology transfer. Sadly, the latter expectation is usually a vain one unless a country already has a strong technology base of its own.

That has not deterred most of the world's nations from competing for high-technology industry. Just as have the several states, many have set up special economic development agencies designed to bypass much government red tape and deal directly and efficiently with high-technology companies. Some of these such as the Northern Ireland Industrial Development Organization (NIIDO) are virtually autonomous and empowered to deal with new companies on a range of issues far beyond those that most domestic development agencies can handle on their own authority.

Overseas organizations like NIIDO offer as incentives the same range of financial and personnel benefits as do the U.S. organizations. And as do the U.S. states, other development agencies such as the Industrial Development Authority of Ireland (IDA), the French, German, Italian, Danish, Luxembourg, Malaysian, Singapore, Sri Lankan and more will try to match those incentives. Indeed, as agents of sovereign governments, not governmental units, these overseas organizations can compete at an awesome level, one few states could match.

Though location per se is seldom the primary factor for a market-driven facility, it can be important if the market is an export one and the company has extensive overseas interests.

Northern Ireland gives ready access to the European market and particularly to the British and British-dependent markets from within the political boundaries of the United Kingdom. Ireland is an English-speaking jumping off place for all Europe from within the Common Market. Luxembourg is a money center. Austria is the gateway to Central and Eastern Europe and Hungary is the stepping stone.

Caribbean dependencies of European countries, such as the Netherlands Antilles, allow favored access to the Common Market from bases closer to home. Special agreements allow other countries like Trinidad and Tobago such access too. Puerto Rico and the U.S. Virgin Islands confer tax benefits no state can match. A high-technology firm can serve markets in Asia better from Sri Lanka or Malaysia.

Most nations have active investment-promotion programs. In some, in many U.S. states, local units such as Glenrothes in Scotland work in parallel with the national agencies. As in the states, the degree of skill, resources and commitment each nation brings to its promotion effort varies enormously.

Profit is no longer a dirty word in many countries. Foreign investment, once shunned, is now sought. But old attitudes die hard. As one skeptical participant in a recent investment promotion conference said: "Politicians propose; bureaucrats dispose. What have you done about your bureaucrats?" The more successful organizations like NIIDO and IDA were designed to shortcut bureaucratic channels. For lack of that authority, some nations' efforts remain hobbled despite the best intentions.

Numerous investment promotion agencies have branches in the United States. Reflecting their independence, these are not connected to embassies or consulates. Other nations treat development as a sideline for diplomatic personnel.

At the urging of a staff member, Adly abd el Meguid of Egypt, the United Nations Industrial Development Organization (UNIDO) in 1978 set up a program to help governments develop their own investment promotion capacity. The U.S. Overseas Private Investment Corporation (OPIC) helped fund the project. Several countries run their promotion efforts out of UNIDO in New York. Sri Lanka was one of the first and more successful graduates of that program. The country's success reflects location and commitment.

THE REAL NORTHERN IRELAND:

Martin Malone is one reason electronics investors score big in Northern Ireland.

Martin Malone knows his way around both the sports circuit and the printed circuit.

A semi-professional Irish League-footballer in off hours, on the job Martin captains a team of 110 printed-circuit technicians as works manager for Irlandus Circuits Ltd. in Craigavon.

Because our universities and technical colleges turn out a plentiful and welcome supply of specialists in electronics engineering and computer sciences. Northern Ireland is an ideal place for any state-of-the-art company to introduce high technology.

Besides our skilled work force, other good reasons to invest in Northern Ireland are an enviable labor relations record, excellent R&D support programs, top-grade transportation and communications links for easy access to the Common Market and other parts of the world, and investment incentives which have been judged the best overall in the EEC.

If you have any doubts about how smoothly business continues to function today in Northern Ireland, ask anyone who's been here recently. Better yet, come see for yourself. For further information, contact Ian Walters or George Forster at the Industrial Development Board for Northern Ireland, New York office, 150 E. 58th St., New York, NY 10155. Or call (212) 593-2258.

Northern Ireland works.

A visit will convince you.

IDB Northern Ireland

SOLID STATE.

The recent growth of "high-tech" industries has underscored the value of New Jersey's long term commitment to science and technology.

From Princeton's Institute for Advanced Studies and Rutgers, the State University, to the 28 state-approved vocational schools, the goal remains the same — to provide industry with the productive, quality workforce so critical in today's competitive marketplace.

Consider our educational resources, for example. New Jersey has five universities, 21 four-year colleges, 22 two-year colleges, a state college of engineering and architecture, plus a state medical and dental college.

Consider research and development. Nearly 10% of the nation's total R&D dollars are spent in New Jersey. 90% of these funds come from private industry. This makes New Jersey not only the leader in privately funded research, but the nation's leader in R&D overall.

Moreover, the State Government is dedicated to creating ideal conditions for business growth in New Jersey. In 1982, the N.J. Economic Development Authority channeled over $850 million in low interest financing, to encourage new business development and expansion here.

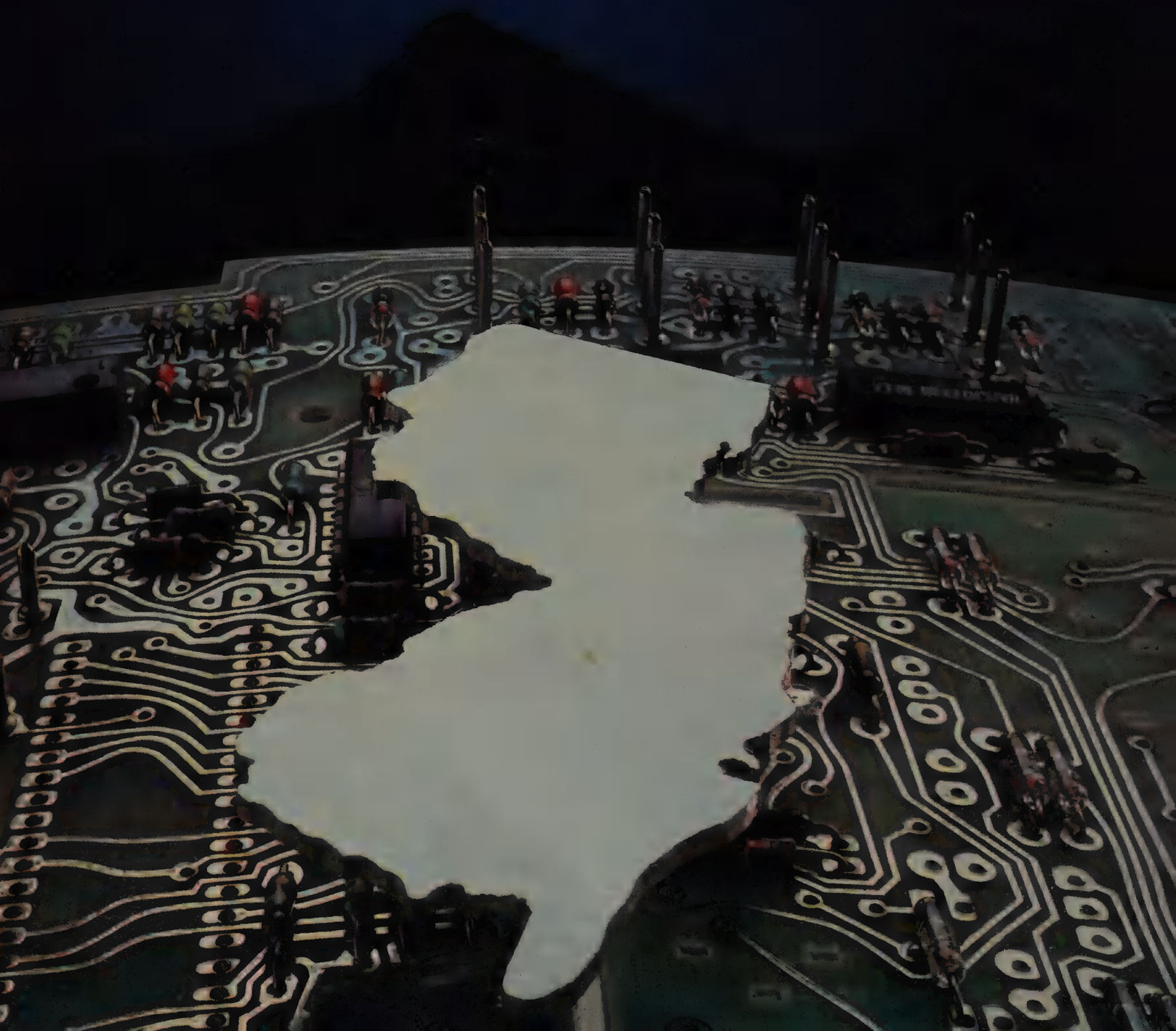
New Jersey also offers many advantages in terms of location, people and lifestyles. The talent of our people is but one reason for saying: "New Jersey and You... Perfect Together."

To learn more about New Jersey contact:

Borden R. Putnam, Commissioner
Department of Commerce & Economic Development
CN 820, Trenton, NJ 08625
Phone (609) 292-2444

Thomas H. Kean, Governor

NJ & You
Perfect Together



BOOKS

Mineral origins of life, the senses, image processing, an ideal monastery, armaments

by Philip Morrison

GENETIC TAKEOVER AND THE MINERAL ORIGINS OF LIFE, by A. G. Cairns-Smith. Cambridge University Press ($29.95). Once upon a time there were those who held that a self-duplicating system might in principle lie beyond scientific understanding. A machine that could build a simple device, say a pair of scissors, is not too complex to conceive. If the machine were asked to build more complex mechanisms, however, it would need to be more complex than they are. A machine able to build itself, which plainly would be no simple mechanism, might then be a logical impossibility, a vicious regress into infinity.

It was John von Neumann, engaged by what he had newly come to expect of the digital computer, who before 1948 brilliantly countered that rather vague argument. The core of his account bears repeating: It is not hard to imagine a machine able to correctly assemble any finite set of parts drawn from a large store. The machine is complicated, to be sure, but it can be assembled out of the same parts. It builds by following taped instructions. Call it a universal tool machine, to which it is an analogue of sorts. Now code the recipe for the tool machine itself onto a long tape. Such a tape and tool machine can then together self-reproduce without logical troubles, given a few more items to allow copying and the like. The tape has taken over all the instructional complexity; it will be long indeed, but the tool machine has reached its plateau of complexity once it is able to carry out an arbitrary sequence of taped orders. It has only to work away for a long time.

We now recognize the profundity of von Neumann's logical example. It describes all life. What is the living cell but a long tape of DNA informing a tool machine? It is all constructed by linking rather simple monomers into long sequences that fold themselves into the molecular structures and catalysts of life. Replication proceeds exactly as envisioned, with the tape calling up a new tape and a new machine out of an existing machine that studiously scans the ancestral tape.

High on the agenda of molecular biology is an account of the origins of such a system. The problem stands unsolved. It is clear that some small length of early tape needs to call up a working tool, to make up another little segment of tape and so on. Here trouble comes. It is the very separation of DNA and the functioning of protein that lies at the heart of the von Neumann logic, and just that distinction makes an initial coupling hard to find. All the wonderful codes and adapters are fine once they are evolved but look implausible as early elements.

This innovative book by a Glasgow chemist expands on an idea he first put forward in 1966. Although the book (which is accessible, he says in his preface, "if you enjoy reading, say, *Scientific American*") does not mention the von Neumann argument, it flourishes within a gap in it, a gap not so much in logic as in the implications drawn from the strict original proposition. The universal tool machine is a nice contrivance if you expect to oversee the evolution of that intricately woven fabric of diversity that now is life on the earth. In the beginning, however, no such flexibility and versatility need be provided. It is enough to have some crude self-replicator, if only there is future opportunity of passing over to a von Neumann scheme, with its unendingly evolving recipes for varied actions that remain chemically distinct. At the outset no such clean distinction need have existed; the plan and the operating devices might have been much more alike.

The primary organisms might have been fabricated at a much coarser level. Today the DNA stores key instructions, not in one or two atoms per bit, presumably the limit of fine-grained instruction, but in molecular aggregates of hundreds of atoms. Early in the course of evolution the miniaturization may have been much less advanced. Today the organic monomers of life are not found outside the biosphere; originally the material basis of information storage for replication may have been inorganic, mineral, abundant in the earth before any life quickened. Today the key sequences are stored in a fundamentally one-dimensional array, the double helix of DNA, wonderfully spooled and reeled in the chromosome dance. In former days the store might have been two-dimensional, some defect pattern spread across a crystal surface layer, able to promote its own reproduction, albeit only approximately.

Dr. Cairns-Smith starts out with a few chapters in the vein of these arguments: definitions, general questions, even a pleasant dialogue in which he concedes that most of his colleagues in the field see his proposals as having "all the qualities of a bad theory"—easy and inviting reading. About half of the book explains the most unfamiliar and positive part of these ideas, the nature of clays, microcrystalline material that forms out of the weathering of common rocks in water solution. The pictures here are as instructive as the text; the fibers and wormlike lengths of many-layered silicate microcrystals offer a very different view of crystal growth. The point is made that in this domain it does not seem unlikely that defects or even accidental linkings of crystallites might confer a selective growth advantage. A simple scenario, taken as an example only, imagines special clays that might "swim upstream to lay their eggs." It is not absurd; consider that a particular configuration might grow well in some interstice in a wetted sandstone. It might turn flocculent after growth, to drift off and seed itself downstream normally. Downstream, however, the environment might change for the worse. Perhaps some of the clay complexes become flocculent once the concentration of sodium ions drops; those individuals break off only at that time, a clear sign of backflow, to seed more survivably in fresher waters upstream. Given self-reproduction, this seems a plausible example of how marginal natural selection might work on colloids.

After many ingenuities, particularly around the ways in which folding and twinning and surface contact might give rise to broadly inheritable surface pattern among crystallites, the book closes with a tale of revolution. Carbon atoms took over and in the end cut loose from their silicate substrate to establish cells. This part of the story is still more conjectural; it calls for detailed biochemistry. It nonetheless has one strength: the membranes of the living cell are a good deal more evocative of a legacy where they are coupled to a layer of clay than they are of any floating loop of DNA.

Right or wrong, this audacious, informal, instructive book is a scientific thriller. It ends in challenge: the answer to these questions will be given not by tighter argument but by finding out how to make some kinds of primary organisms in the laboratory. That result is not

close, but it can be foreseen; the first hints of surface pattern reproduction in clays may already have been found, as long ago DNA was shown to replicate in the test tube. The parallel is cautionary; this is no fleeting puzzle but a riddle grander than the Sphinx ever put.

THE SENSES, edited by H. B. Barlow and J. D. Mollon. Cambridge University Press ($59.50). That the human senses are to be seen as physical instruments is an insight as familiar as it is profound, and the opening chapter of this compact and up-to-date summary directly applies the metaphor. Plainly it makes a difference whether the laboratory example is the camera obscura of the 17th-century philosophers or the digital computer with video input of today. Fourier synthesis and signal-to-noise ratios inform many chapters; they are elucidated graphically and well without formal mathematics. One experimental result put into this context speaks eloquently of unity. A histogram plots the number of nerve impulses recorded at synapses in the spinal cord but originating in a single sensory terminal in the skin. No stimulus was applied: the noisy nerve fiber fired anyhow, with a mean rate of six impulses per tenth of a second. A moderate stimulus doubled the rate of firing. In the two cases the observed firing rate fluctuated at random by about the same amount, plus or minus three pulses per tenth of a second. Thus both the signal and the signal-to-noise ratio doubled. The error rate for the stimulus would be about one in six times; for a judgment to be made at odds of one in 100 the signal would need to be twice again as strong. This is neuroengineering.

A recent experiment displays the response of a single retinal rod cell (it is one taken from a toad retina, an order of magnitude longer than a mammalian rod) to single photons. The current flow to the rod caused by the absorption of light by rhodopsin was recorded for 40 weak flashes. At the time of the flash the response duly appears, except that many bumps are missing. The Poisson statistics hold well because the events are single-quantum ones. Our own color-detecting retinal cones, much shorter and less absorbing than the color-blind rods, need many photons if significant current is to be seen above the background noise. Now, cone color response is shifted about 10 percent redward of that of the more sensitive rods; it is hard to resist the inference, although it is far from proved, that the higher thermal noise faced in the cones because of their lowered energy barrier has called for a reduction in sensitivity. Rods are fit for detecting faint images at the periphery; cones, for acute study when there are plenty of photons. The design of the retinal surface is understandable.

The eye is also, of course, a classical lens system; the most frequent clinical intervention is for lens improvement. The refractive model of the eye is here, and with it one sees a strict measure of performance. The measured spread of a bright-line image is not much broader than the diffraction-limited theoretical best to be had with a given pupil size and color of light. The eye, however, is no mere camera, not even one with a remarkable dual film response: the "sensitive long-latency, slow-exposure monochromatic system (the rods) with a less sensitive, shorter-latency, brief-exposure color system (the cones)," both speeding up automatically as the light level rises. The eye perpetually scans. That foveal spot offers five times the peripheral resolution over a central region only a few degrees across. We carry a detailed survey across the entire visual field quite unconsciously in fifth-of-a-second hops, neither aware of the motion nor bothered by the piecewise detail in the scene.

Such subtleties, along with stereo vi-

sion, edge and feature detection, and color perception constant under wild changes in the color of natural illumination, belong to higher centers. Some lie in the retinal interconnections, some close to the cortex itself. There is the computer hard at work in us all, open to laboratory study only since that metaphor itself became realized within the past generation.

About half of this unusually well-written and well-coordinated account by a dozen British research workers (plus one Michigander) is spent on the eye. Another fourth of the volume deals with hearing, presenting a good mix of the familiar with the brand-new. Nothing is said explicitly of hints of regenerative circuitry in the cochlea, although the unexpectedly sharp tuning of that wet system draws attention.

Last and least, although perhaps freshest, is the treatment of the other senses: touch in skin and viscera, the system of the inner ear that is our version of inertial guidance, the chemical channels of smell and taste. There is good evidence of a descending control outward from the brain along the pathways usual for inflowing pain. Biochemical study is rapidly identifying specific chemical mediators and antagonists in service throughout this intricate alarm system.

One remarkable result almost a decade old is the induction of vertigo by drinking heavy water. The interpretation is that the stuff entering the fluid of the semicircular canals increases the fluid density. The slight mismatch in buoyancy then makes the delicately wired diaphragm that blocks the canal sensitive to its position under gravity instead of to motion alone, the normal state. A stagger results, akin to that proverbial after unwise intake of alcohol but with the apparent turning being in the opposite direction. Alcohol may act merely to lessen the density of the canal fluids. The result seems almost too pat to be true, even though it does account for such a specific effect from such a simple substance as ethyl alcohol.

Throughout the mind's instruction is under study with oscilloscopes and signal generators; issues of nature and nurture, of perception and cognition arise almost as naturally in the laboratory as in the physiological details cited. The topic of this contemporary, ambitious and reflective treatment transcends medicine.

DIGITAL IMAGE PROCESSING: A SYSTEMS APPROACH, by William B. Green. Van Nostrand Reinhold Company Inc. ($34.50). It is nothing new to regard an image as a large array of numbers. But this book, quite accessible to a general reader with a limited tolerance for algebraic equations, provides an up-to-date overview of the consequences of regarding that array as an invitation to manipulate. First the artists, then the cartographers, then the photographers and now the computer engineers have used the image as raw material. Of course, eye and brain have always done no less, so to speak in secret, but it takes the digital computer, nothing if not explicit, to make clear how much everything we see is a piece of applied epistemology.

Two fine pictures of crescent Mars face each other, products of the cameras of fast-approaching *Viking Orbiter 2*. One shows the Martian polar cap white below a beige quarter planet mottled with detail. The computer has already been at work. As the spacecraft moved in, three successive views were geometrically registered in spite of steady changes in size and aspect. The accuracy of fit is about as good as possible, to a single picture element, or statistically even better. The technique is powerful; several small features common to all the images are chosen as reference points. The statistical correlation between the numerical images is then computed for

a spread of trial alignments in the approximate neighborhood of the recognizable feature. The highest correlation is chosen; repeating the procedure for several tie points, a transformation can be numerically defined to register the distinct images sometimes with "subpixel alignment."

Each of the three Martian views was shot through a different color filter. The camera response was known to be distorting. It had been calibrated by recording a uniform field at different intensities. The corrections were then applied to each image, to yield a set of three single-color images. These were then extrapolated to provide red, green and blue images with bandwidths matched to the requirement of the color film.

So far the task has been remedial. Now for enhancement. The Vidicon camera has collected close to a million numbers per image, each number in the eight-bit range from 1 to 256. The computer takes a census. The head count by intensity of pixel in each of the three colors is graphed right on the margin of the picture. These image histograms are basic tools for the work. The three bland profiles, however, do not mean much; we can see no easy outcome. We need a theory. It is the standard theory of three-color reproduction. The three intensities at one pixel—red, green and blue—define a point in the familiar curvilinear color triangle and also a total intensity.

Now the color space can be represented anew. The three coordinates yield one number to represent intensity, a second to represent the hue (regarded as an angle turning from the green apex around through the other primaries) and a third to represent saturation, increasing with distance out from the white point within the color triangle. These three numbers can be computed for each pixel, and again a census is reported. One stretches the intensity distribution in a simple linear way, to take full advantage of the film latitude, in this case no large effect. The hue and saturation populations are left unmodified.

The result is "the best true-color rendition of the planet that can be achieved" with the system: not much color. The facing page shows the same scene, the same size, the same features. A watercolorist, we imagine, has been at work. Subtle reds and even some greenish tints have been added tastefully to enhance the interest of the beige ball. That was not, of course, the origin of the comparison image. It is a controlled, objective product, even though it looks like a painting. Its hue distribution was spread out by a factor of seven and its saturations by a factor of three, holding their average values fixed. This is the Martian surface seen by an eye of superhuman abilities; color nuances have become bold.

Most of the technical volumes in this field concentrate on the mathematical aspects of the algorithms that can economically carry out such schemes on pixels by the million. William Green has instead chosen to tell this tale in the large, emphasizing the chain of operations, the techniques in general, the choices and the tradeoffs. Illustrative comparisons such as the one outlined above abound, and with them the block diagrams of the systems required. Green summarizes the novel camera-equivalent devices—in satellites, in the medical center and in astronomy—that acquire the pictures, the schemes by which the pictures can be digitized, the growing technology of easy display and access to images in enormous numbers, and the all-important details of their management, such as careful labeling and filing schemes of high flexibility. Images so elaborately treated are misleading without their clinical history, unless long familiarity and "ground truth" are at hand to provide the interpretation out of experience itself.

Many other paths to enhancement of image are described, including more subjective stretches of iterated contrast and the simple color substitutions that are now familiar, particularly in the pages of *Scientific American*. There are new experiments with thematic maps, where a space is defined, say, by the colors of known crops and each pixel is given a coded color in accordance with some measure of its position in the color space. The result looks like a colorful map of remarkable detail, and it would not mislead a canny viewer as the processed photograph can. We recognize at once that maps are partial representations based on understandings, that is, on theories, although we have not quite learned that no image at all, not even the field of view of every waking moment, is in fact much different.

The references here include books and the periodical literature, as is usual. They extend, however, to rich sources of images of earth and sky and to sources of software for image processing, available from a NASA agency at the University of Georgia called COSMIC. Green, a senior figure in the field at no advanced age, "was fortunate to spend ten years at [the Jet Propulsion Laboratory] working in the finest image processing facility in the world." The examples reproduced in the book support his claim, although we may wonder about the inaccessible contractors to the National Reconnaissance Office.

THE PLAN OF ST. GALL: IN BRIEF, by Lorna Price. University of California Press ($55). Some three years ago there appeared a work of intellectual and visual reconstruction as beautiful as it was learned, a "study of the architecture and economy of, and life in, a paradigmatic Carolingian monastery." Its devoted authors, Walter Horn and Ernest Born, had spent some 15 years at the task, and they were able to express their findings in a form worthy of the scholarship, a classic of our day, acclaimed for its encyclopedic richness no less than for its perfection as a printed and illustrated book.

At the same page size, but only a tenth as thick, this new book presents (at roughly a fifth the price) a sample and a summary of the parent work, all in vermilion, black and parchment. The author was the editor of the main volumes, and her explanatory text is set in a design by Professor Born.

What is here is perhaps best described as a chain of projections linked across 11 centuries. In about A.D. 820, when Louis the Pious held the throne of his father Charlemagne, a wonderful plan was made in his palace at Aachen for the ideal structure of a monastery of some 300 souls devoted, under the rule of St. Benedict, to the work of God. It honored and framed the work of the hands—and of hens and oxen and water mill too—that jointly sustain the Opus Dei, under abbot and king. That original plan is lost; what we hold is a single work of art, done in red on a pieced-together parchment almost a meter square, devotedly traced from the original by two men who worked in the scriptorium of Reichenau before A.D. 830. Since then the copy has been held in the monastic library of St. Gall, above Lake Constance. At about the end of the 12th century an unknown man there wrote on the blank side his pious life of St. Martin; he must have admired the old plan, because he was careful not to cut the parchment, although he erased one of the 40 ground plans it bore.

Then the document vanished, although it was still safe on the shelf because it was catalogued as a life of the saint. In 1604 it was returned to the stream of knowledge, its hundreds of detailed legends published without any drawings. In the mid-19th century scholarship touched the plan of St. Gall, but not until 1952 was a facsimile published. As a part of the observation of the 1,200th anniversary of the Carolingian reign the two Berkeley scholars were charged with a study of the plan. Their massive work emerged in time to universal praise; this, its "overview," preserves beautifully a terse narrative and many illustrations prepared for an exhibit of their findings, which will tour a dozen American galleries between 1983 and 1985.

Two major tasks sprang from the plan: to project a third dimension out of each of the careful ground plans neatly composed in elegant rectangularity and to discover the size and scale the plan-

ners had intended. The shelters, gardens, bathhouses, cloisters, infirmaries and cemeteries were given full form by the intensive use of comparisons, often archaeological ones. The Church of the Plan, never built, is rendered in quiet nobility, along with its cloister and refectory. The functional structures were not modeled on the classical; indeed, they "descend from a house type" that was common among the Frankish people. The present St. Gall monastery is only of minor help, because there "no stone of Carolingian origin stands today above ground level." Fires, enlargements and the sacking of the site in 1712 have disturbed the signal in the stone. In fact, although Abbot Gozbert depended on the idealized plan as a guide to his A.D. 830 rebuilding of the monastery, then rather seedy, the hilly site could not fully accommodate the paradigm; the adjustments were major.

Searching out the scale revealed a surprise: the design was governed by an unexpectedly harmonious system. The architects followed a rigorous "progressive dichotomy," the modules they employed spanning no less than eight powers of two from great nave to burial plot. Among all these consistent four-square relations some circles appear: the two apses of the church are semicircles, and its two high detached circular towers, without bells, are imposing announcements of "a Fortress of God." A couple of 40-foot circular hen houses, an "exemplary solution" for a busy community that daily needed the eggs of hundreds of birds, are the least expected element of this vivifying display, an extraordinary medieval union of the workaday and the exalted.

WHAT KINDS OF GUNS ARE THEY BUYING FOR YOUR BUTTER? A BEGINNER'S GUIDE TO DEFENSE, WEAPONRY, AND MILITARY SPENDING, by Sheila Tobias, Peter Goudinoff, Stefan Leader and Shelah Leader. William Morrow and Company, Inc. ($15.95). HOW TO MAKE WAR: A COMPREHENSIVE GUIDE TO MODERN WARFARE, by James F. Dunnigan. William Morrow and Company, Inc. ($14.50). In the guns-or-butter title of this pair an academic team has assembled a well-written nonmathematical primer of the nature of modern war, both nuclear and conventional, with such matter-of-fact grimnesses as the following. Old, densely built cities, particularly if there is much wood in them, support fire storms, given a large enough initiating source. That is what Hamburg, Dresden and Tokyo shared, to say nothing of Hiroshima. American suburbs, however, are built with material an order of magnitude less combustible, only two pounds per square foot; they will not sustain a fire storm. Nevertheless, "older American cities, [such as] Baltimore and Philadelphia, would surely burn."

The bulk of the argument is presented in historical context, with a chapter that starts with sticks and slings, another that carefully traces the rise of nuclear weapons and their strategies, and another on efforts toward arms control. A long chapter treats of today's conventional weapons, from rifles to aircraft carriers. The text is narrative and informal, and the authors have quite successfully managed to dispel jargon, although the reader will learn what an M-16 rifle and a P-3 marine patrol aircraft do (and look like). Simple description fills the symbols with content, but it is not easy to talk about such organisms without their names. (An unusually "long aside" takes about seven rather lighthearted pages to give some rationale for Pentagonal nomenclature, and there is a longer annotated vocabulary of these terms of art as well.)

If half of the book is military history and technology, half is rather cool politics: structure. The defense experts, the military, the Congress, the defense contractors great and small (the great ones are listed) are paraded in review, both as individuals and as classes, in a persuasive chronicle of how the fast-growing 60 billions or so spent on weapons procurement annually (the figure is up again since the book was written in 1982) are claimed and used, in "the single most complex process by which policy is made in our society."

A similar tale is told for arms control—a shorter story—and the book ends with a look at issues of the decade: NATO, a bigger navy, the Rapid Deployment Force, the embattled MX, high-tech v. low, and more. The volume shares the breezy tone of its title, but it is in fact an extensive and serious answer, evenhanded overall, although the authors are plainly engaged and here and there show a claw, if not a tooth. The Mine Warfare Command, they reflect, is starved for skilled people. "It lacks glamour. You simply plant it in the water and go away. It does its work in your absence. No glory there." Over a wide span of material the book is accurate; here and there a technical misunderstanding creeps in, as it does for the beginnings of thermonuclear weapons in the two superpowers. This team of authors nonetheless knows how to inform serious beginners. "We don't want to tell you what to think," they write, "but what to think about."

How to Make War, the second of these books, is again an overview of war and weapons today, with a title that carries the full flavor of the somewhat ironic text. James Dunnigan's vantage is a unique one. He makes war, professionally, not in real life but in simulation, as the author and publisher of games and simulations of combat and campaigns through history. His aim is an appraisal of the means and nature of war today, and he treats explicitly those high adversaries, the U.S. and the U.S.S.R., with the other nations at least sketched in.

The book is really an enfolding of two works, both first-rate for readers with some experience. About a third of it consists of a long series of tables of data (typographically hard to read) along with terse, defining glosses and enlargement. The dozen-odd tables include the expected lists of the characteristics of, say, battle tanks (the Russian tanks are in "massive numbers . . . effective yet expendable") or strategic weapons, complete with kill factors. They include as well the supply needs of divisions and of aircraft units, the estimated combat losses according to branch of service, the ground density of personnel in units of the American and Russian forces, and the itemized costs of raising an infantry division. These tables surely derive from the handbooks available for the war games of the author. They are not documented or even dated; they appear to be of the epoch 1980, conscientiously compiled from annuals and field manuals.

The remaining text is an informal, personal ("Imagine yourself in a suburban town that has been abandoned . . .") set of brief background essays on war by land, air and sea, on war with special weapons, on logistics and transport and on the human factors less easy to enumerate. The essays read well; the informality and the point of view are reminiscent of the reflective talk of soldiers. They do not neglect the dilemmas of the discipline of making war. Technical analysis and the lessons of history alike claim audience, although they frequently conflict. Only a few pages are spent on the choices of lawmakers and citizens; the book views a soldier's world, with the forces, enemies and allies given, but it allows for change.

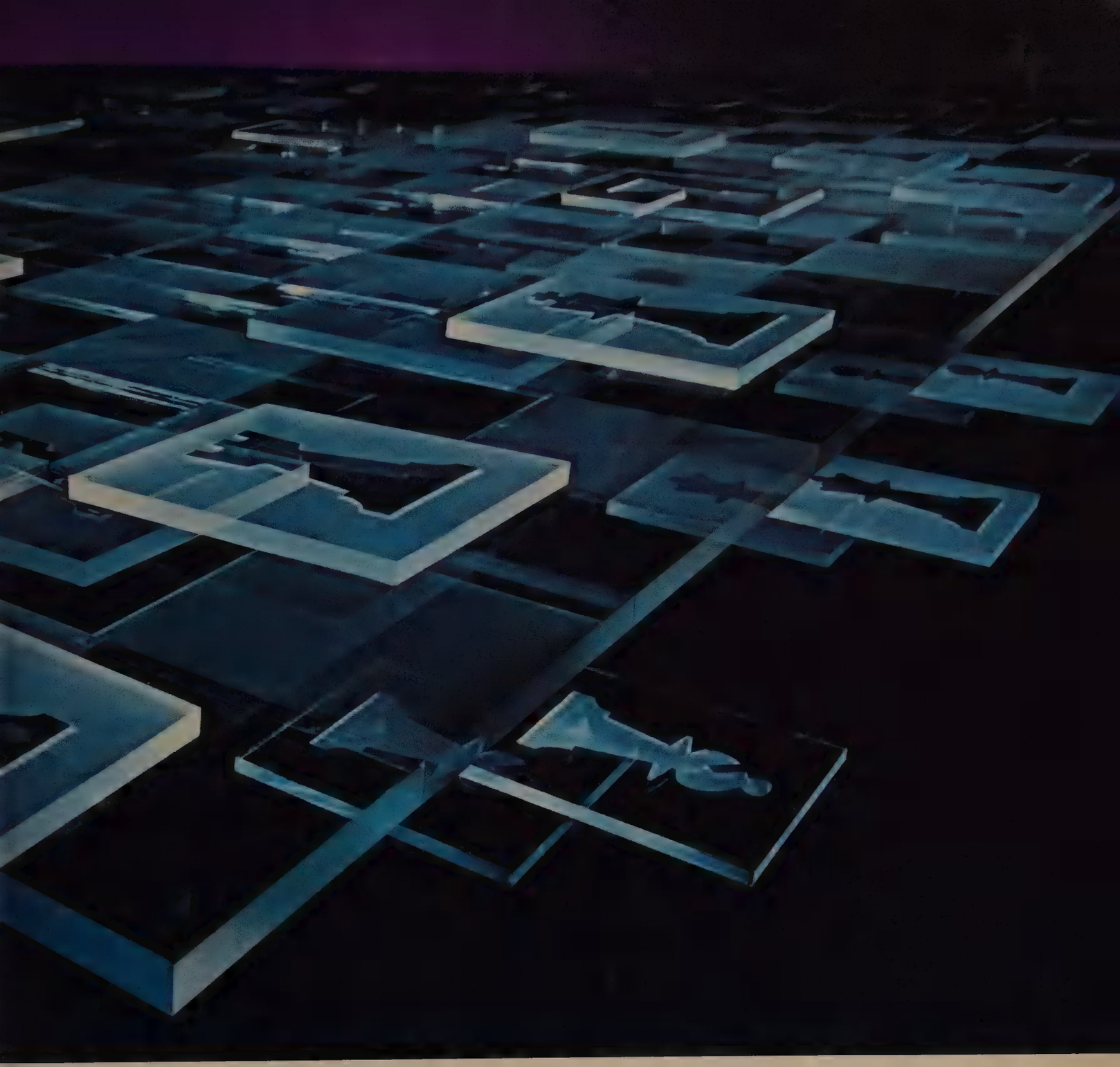
Out of all this observation a few conclusions are presented rather as dry aphorisms, although they are well if implicitly argued. One year of a major conventional war as now planned would cost the U.S. more than $1 trillion. Most of this is the cost of air war above land and sea, of which most is the cost of current "smart" munitions. Production would need to rise thirtyfold over a couple of years to make it all possible. A short war is the planners' hope—perhaps the planners' myth. On naval warfare: "The only thing that will defeat the West at sea is complacency or bankruptcy from building too many ships." On nuclear doctrine: "Both Russia and the United States have about ten times more nuclear warheads than they need to guarantee mutual destruction. In such a situation, nuclear 'tactics' becomes irrelevant."

An IBM 4341 super-mini: chip design at Hughes

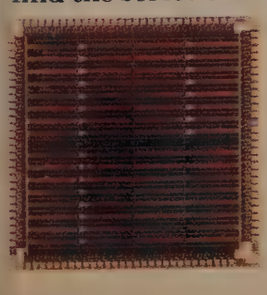
The board's half the size of a thumbnail, with 100,000 pieces. And it's your move.

Finding paths for new interconnections on an integrated circuit is like no chess game you have ever seen. The first few moves are easy, but with up to 100,000 electronic devices crammed onto a minuscule silicon chip, the possibilities soon become astronomical. That's why engineers at Hughes Solid State Products in Newport Beach, California use an IBM 4341 super-mini.

Brian Tien, head of design automation, says, "Without the IBM 4341 we couldn't get this much function on a chip. With it, we can finish a circuit in a few weeks instead of many months."



The engineers work interactively with the system, assigning logical functions to devices on the chip. Then, using software, they route conductors—deposited strips of metal—to connect the logic gates. If a pathway becomes too crowded, another layout is automatically provided. And another. Until they find the series of moves that works.



This Hughes integrated circuit is a high-density gate array with two-level metal interconnections. Actual size is only 0.3 inches square.

They can think about the problem, and not worry about the computer, with microcode-assisted IBM software called VM/CMS. "Our people find VM ideal for interactive computing," Tien says. "As many as 40 design engineers make demanding use of the 4341 at once. Response time is excellent. The full-screen editor speeds up programming. And the executive language is simple, yet powerful."

The 4341 is an excellent example of IBM technological leadership. Internal data paths, and the arithmetic and logic unit, are 64-bits wide, built of 64K-bit chips that IBM has been mass producing since 1978. Real memory goes to 16 megabytes. The multiple, semi-autonomous processing units use high-density, large-scale-integration technology.

IBM offers engineering and scientific users extensive support: consultants, education and access to professionals. Tap into our many years of experience.

For an informative brochure, write Dr. Jack W. Hugus, IBM Engineering and Scientific Marketing, 1133 Westchester Avenue, White Plains, NY 10604.

IBM®

"Water covers three-quarters of the earth's surface. Our Wagoneer covers the rest."

DENNIS CONNER, SKIPPER OF THE 1987 AMERICA'S CUP WINNER

Skipper Dennis Conner and his America's Cup crew may rule the Seven Seas...but nothing covers the land in between like the Jeep® Wagoneer Limited. And that's why Wagoneer has been made the Official Vehicle of the America's Cup defenders.

Like the boats they sail, the Wagoneer was designed to perform to the limit, even under adverse conditions. They count on the security of Wagoneer's remarkable Selec-Trac 2-wheel/4-wheel drive system - especially when the dry land is anything but dry.

It's also a comfort to know that in a Wagoneer Limited the classic luxuries are standard. From its genuine leather upholstery to its premium sound system, quality is never compromised.

And because it does so many things without compromise, the Jeep Wagoneer Limited has developed a singular reputation as the perfect vehicle for that *other* quarter of the earth's surface.

Jeep® Wagoneer Limited. The Ultimate Wagon.

FROM AMERICAN MOTORS CORP.

Established 1845 SCIENTIFIC AMERICAN May 1983 Volume 248 Number 5

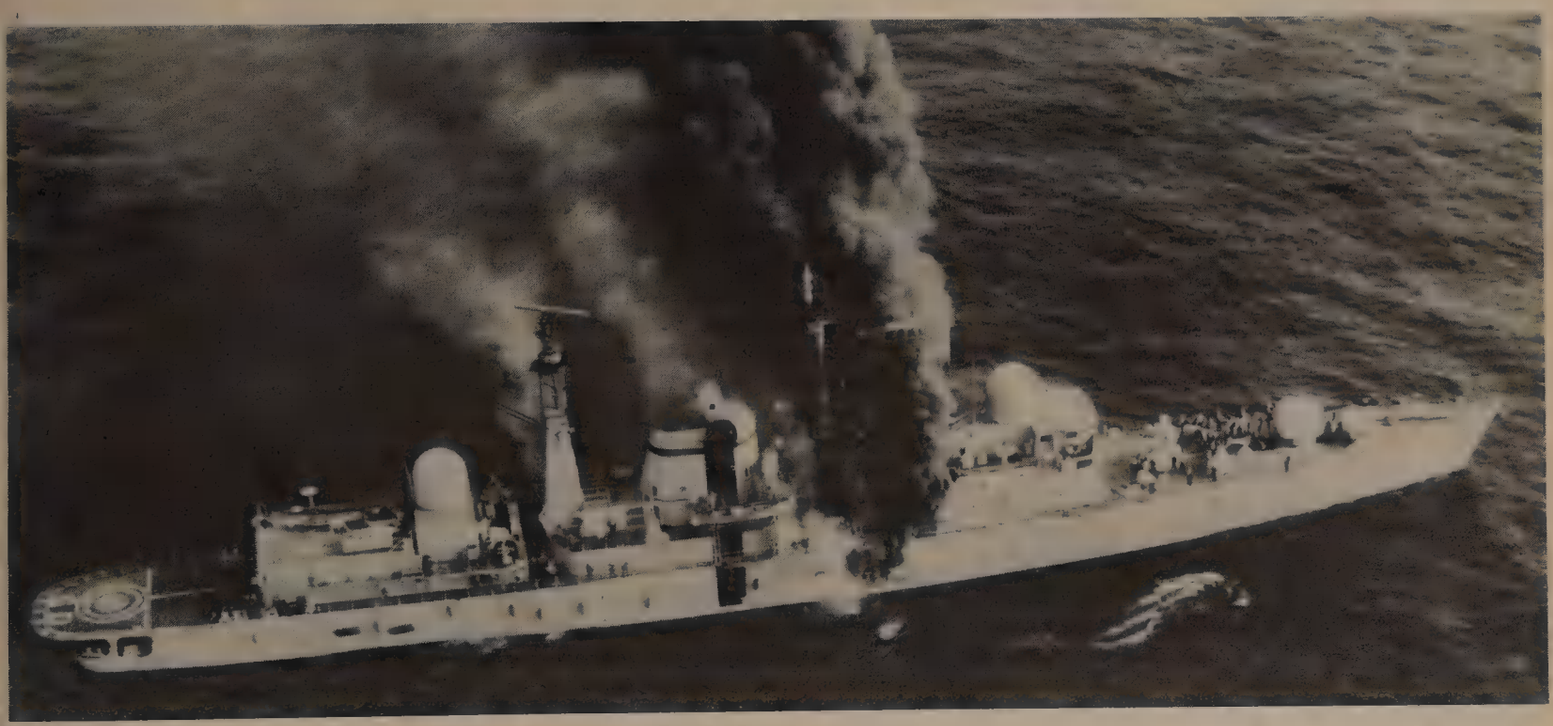
Smart Weapons in Naval Warfare

The impressive performance of precision-guided weapons in the Falkland Islands war calls into question current plans to add more large, expensive surface ships to the U.S. fleet

by Paul F. Walker

Last year's war between Britain and Argentina over control of the Falkland Islands was viewed with particular interest by military planners and others concerned with evaluating the effect of modern precision-guided weapons on the future of naval warfare. Of the 114 aircraft and 10 ships lost by both sides during the two-month conflict more than half were put out of action by such "smart" weapons. Included in the losses were 61 Argentine aircraft shot down by surface-to-air or air-to-air missiles, three British ships destroyed by air-to-ship or surface-to-ship missiles and one Argentine ship (the superannuated heavy cruiser *General Belgrano*) sunk by submarine-launched, acoustic-homing torpedoes. In addition nine British aircraft, 44 Argentine aircraft, three British ships and three Argentine ships were destroyed by more conventional means, such as gunfire, aerial bombing and accidents.

From the viewpoint of naval force planning the most significant loss was that of the British destroyer H.M.S. *Sheffield*. The *Sheffield* was hit shortly after 2:00 P.M. on May 4 by a French-built Exocet missile launched from a distance of about 20 miles by an Argentine navy Super Étendard fighter-bomber (also made in France). At the time of the attack the ship was on picket duty southeast of the Falklands and just west of the main British fleet. The sea-skimming, radar-homing missile struck the *Sheffield* amidships about six feet above the waterline, starting a fire that quickly disabled most of the destroyer's operating systems, including its fire-fighting apparatus. Five hours after the attack,

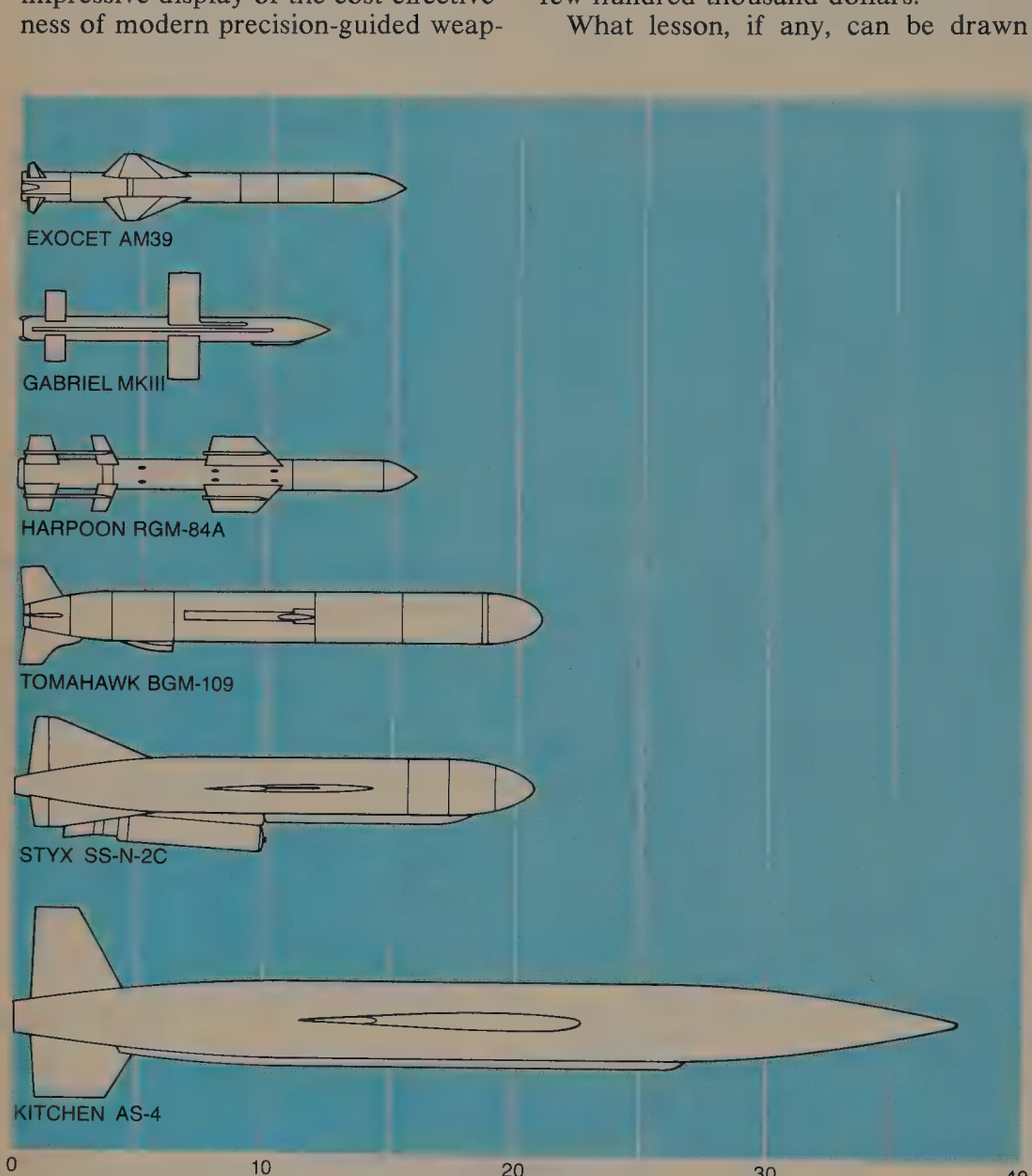


H.M.S. "SHEFFIELD" BURNS UNCONTROLLABLY after being struck by an air-launched Exocet missile during last year's Falkland Islands war. The comparatively small, sea-skimming, radar-homing "smart" missile was fired from a distance of about 20 miles by a low-flying Argentine Super Étendard fighter-bomber. (Both the missile and the aircraft were made in France.) The 4,000-ton British Type 42 destroyer, which was equipped with Sea Dart antiaircraft missiles and other modern fleet-defense systems, was hit amidships about six feet above the waterline. This aerial photograph was made before the order was given to abandon ship. Crew members not involved in fighting the fire can be seen standing on the fore and aft decks. A small inflatable launch moves away from the ship after coming alongside for a close inspection of the damage. The burned-out hulk of the *Sheffield* was later scuttled by the British. Twenty sailors died in the attack.

with the flames spreading dangerously close to the ship's ammunition stores, the order was given to abandon ship. The burned-out hulk was later scuttled after an unsuccessful attempt was made to tow it from the battle scene. Twenty British sailors died in the attack.

The destruction of the *Sheffield* was an impressive display of the cost-effectiveness of modern precision-guided weapons. The *Sheffield* was a highly rated, 4,000-ton destroyer, built about 10 years ago at a cost of approximately $50 million. It was armed with some of the most advanced defensive systems available anywhere today. Yet it was defeated in one stroke by a comparatively small, medium-range missile costing a few hundred thousand dollars.

What lesson, if any, can be drawn from this incident? Does the advent of smart antiship missiles such as the Exocet signal the end of an era in naval warfare, as some have suggested? Are large, expensive surface warships now obsolete—easy targets for a new generation of cheap but deadly precision-guided weapons fired from aircraft, small patrol boats, submarines or shore batteries? Or, as others have argued, are there equally sophisticated defensive measures in the offing that could make such an incident the exception rather than the rule? A review of the technological basis of this debate could not be more timely, in view of the current plans of the Reagan Administration to enlarge the U.S. fleet by more than 100 new warships, including two nuclear-powered aircraft carriers projected to cost more than $3.5 billion each.



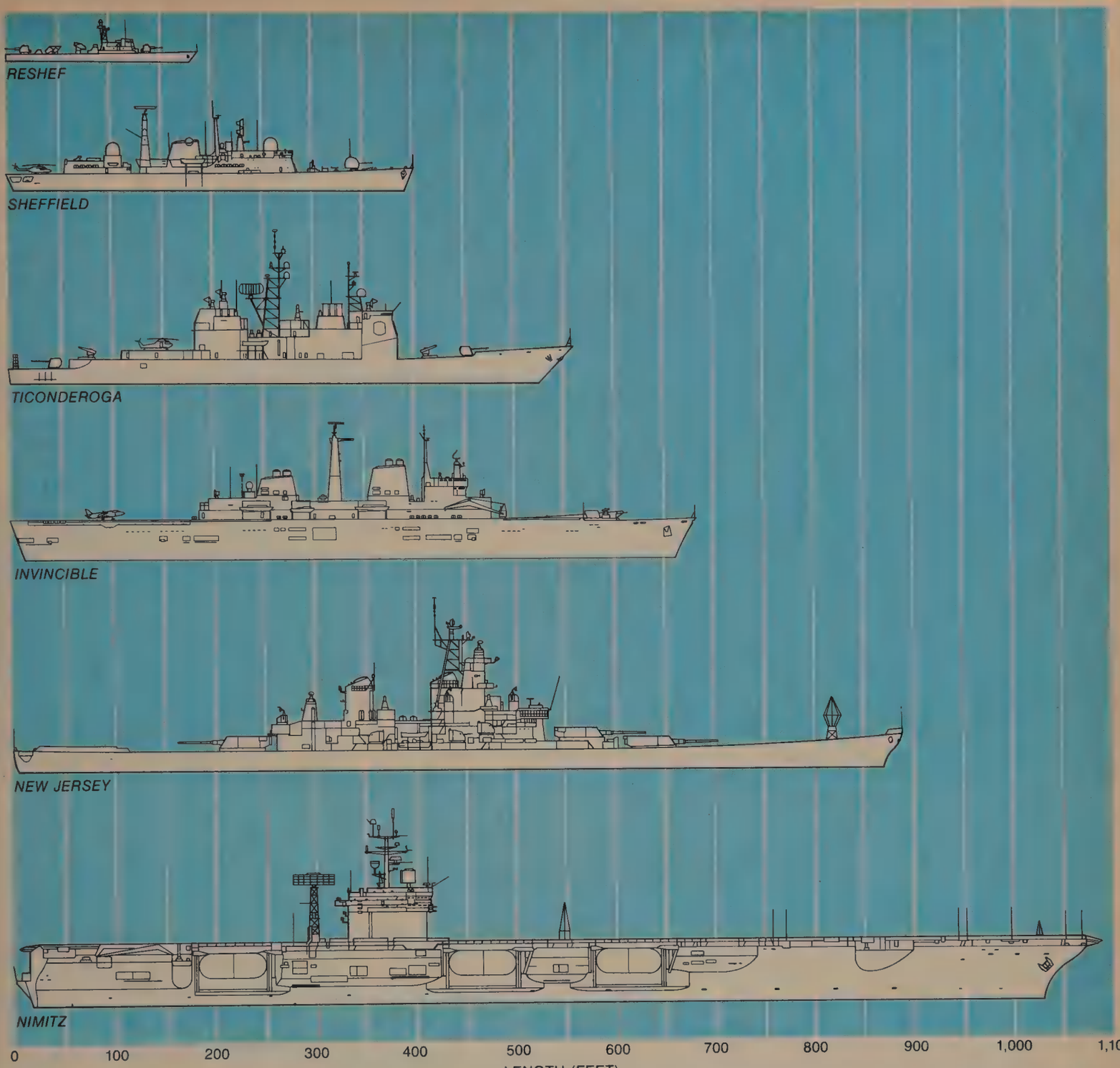
ANTISHIP MISSILE	YEAR DEPLOYED	RANGE (KILOMETERS)	SPEED (MACH NUMBER)	TYPE OF WARHEAD	ESTIMATED COST (1982 DOLLARS)
EXOCET AM 39 (FRANCE)	1979	50+	.93	NON-NUCLEAR	?
GABRIEL MKIII (ISRAEL)	1979	60+	.73	NON-NUCLEAR	$400,000
HARPOON RGM-84A (U.S.)	1977	90	.90	NUCLEAR OR NON-NUCLEAR	$1 MILLION
TOMAHAWK BGM-109 (U.S.)	1982	450	.74	NUCLEAR OR NON-NUCLEAR	$3 MILLION
STYX SS-N-2C (U.S.S.R.)	1965(?)	80	.90	NON-NUCLEAR	?
KITCHEN AS-4 (U.S.S.R.)	1962	300+	2+	NUCLEAR OR NON-NUCLEAR	?

REPRESENTATIVE ANTISHIP MISSILES are drawn to scale in the illustration at the top and are characterized in the table at the bottom. Missiles such as the Exocet are designed to be launched from airplanes, helicopters, ships and land-based installations; the particular model shown here, the air-launched Exocet AM 39, is the one that destroyed the *Sheffield*. The Russian ship-to-ship missile referred to by Western intelligence sources as the Styx was used by the Egyptian navy in 1967 to sink the Israeli destroyer *Elath*. The large Russian missile designated AS-4 is carried by the Backfire bomber; it can be armed with either a conventional high-explosive warhead or a nuclear warhead. Mach numbers are multiples of speed of sound.

The Exocet missile is described by its manufacturer, the French company Société Nationale Industrielle Aérospatiale, as an extremely reliable weapon, "capable of destroying any type of surface vessel" with a "kill probability" of more than 90 percent. Compared with other missiles of its type the Exocet is not very big; it is about 15 feet long and weighs 1,500 pounds, including its 350-pound high-explosive warhead and a full load of fuel. The missile is powered by a two-stage solid-fuel rocket and has a maximum range of between 25 and 40 miles, depending on how it is launched. Different models are designed to be launched from airplanes, helicopters, ships and land-based installations. The Exocet was introduced in 1972 and has been operational in the French navy since 1978; it has also been acquired by at least 17 other nations. So far it has been mounted on more than 180 ships, including the British navy's newest frigate, H.M.S. *Beaver* (which was not ready in time for the Falklands war).

In a typical air launch of a smart antiship missile such as the Exocet the pilot of the attacking aircraft flies low over the water under the defender's radar coverage until it is time for the missile to be launched. The aircraft then "pops up" briefly to get a bearing on the target, dropping below radar coverage again as the coordinates of the target are fixed in the missile's microelectronic memory. The missile is released in a "fire and forget" mode, flying the last leg of its journey under its own power and for the most part under the control of its own inertial-guidance system and radio altimeter; the latter device enables the missile to skim the surface of the water at a height of a few feet. As the missile approaches the target it rises slightly to scan the horizon and lock its terminal homing radar onto the target. The missile's warhead is designed to first penetrate a ship's hull and then detonate.

The Exocet's target in this case, the



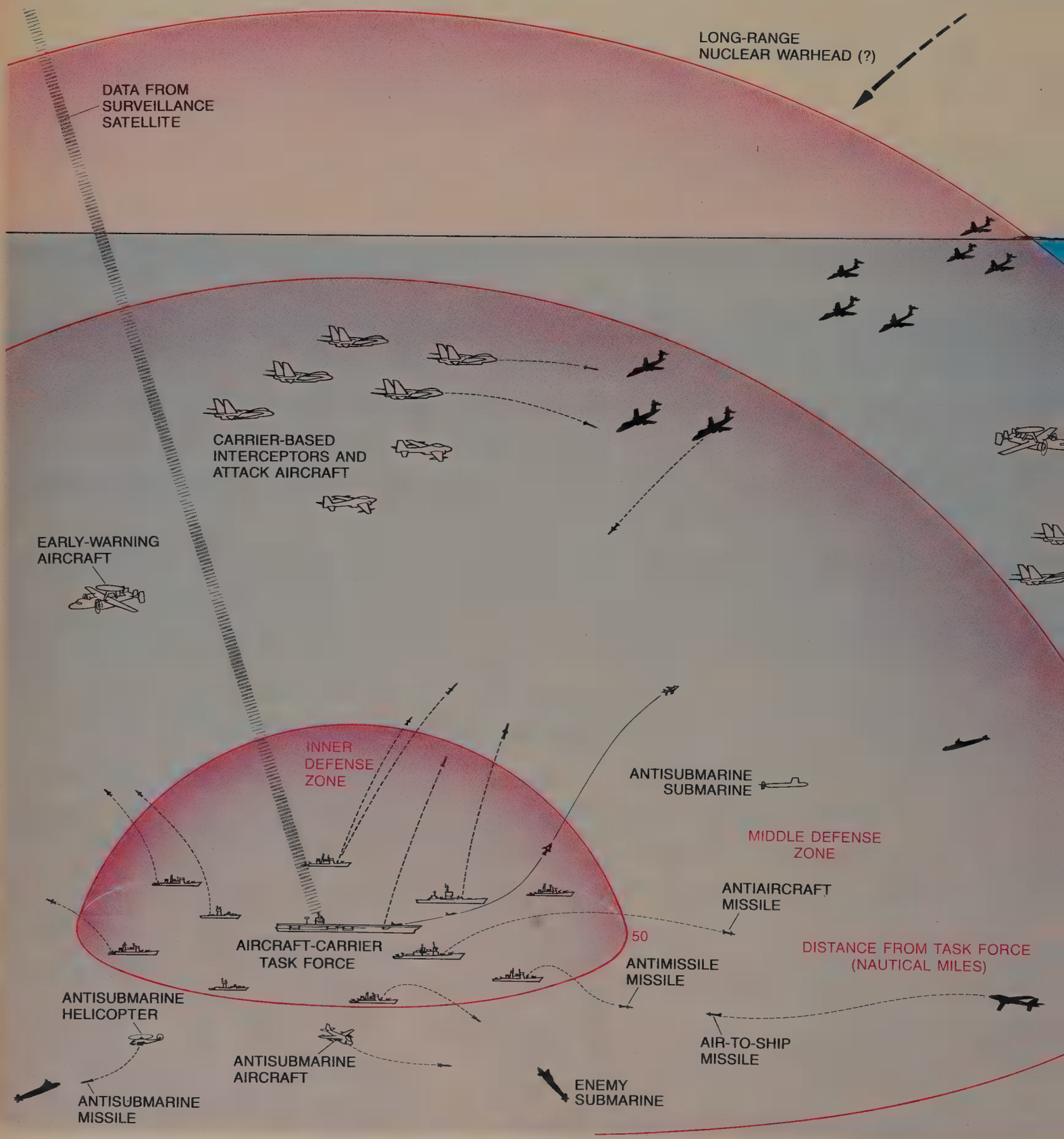
SHIP	YEAR DEPLOYED	DISPLACEMENT (TONS)	CREW	GUNS	MISSILES	AIRCRAFT	ESTIMATED COST (1982 DOLLARS)
RESHEF (ISRAEL)	1973	450	45	2 76 MM. 2 20 MM.	5 GABRIEL 4 HARPOON	0	$50 MILLION
SHEFFIELD (BRITAIN)	1975	4,000	270	1 115 MM. 2 20 MM.	20 SEA DART	1 HELICOPTER	$150 MILLION
TICONDEROGA (U.S.)	1983	9,000	330	2 127 MM. 2 20 MM.	68 STANDARD 8 HARPOON 20 ASROC	2 HELICOPTERS	$1 BILLION
INVINCIBLE (BRITAIN)	1980	20,000	1,300	0	22 SEA DART	5+ AIRPLANES 9 HELICOPTERS	$525 MILLION
NEW JERSEY (U.S.)	1943 (REACTIVATED 1983)	58,000	1,600	9 406 MM. 12 127 MM. 4 20 MM.	16 HARPOON 32 TOMAHAWK	4 HELICOPTERS	$326 MILLION (MODERNIZATION ONLY)
NIMITZ (U.S.)	1975	91,000	6,300	3 20 MM.	24 SEA SPARROW	80+ AIRPLANES 10+ HELICOPTERS	$3.5 BILLION

REPRESENTATIVE WARSHIPS are depicted at the top and characterized at the bottom. The Israeli fast-attack craft of the *Reshef* class are typical of the hundreds of small missile-launching vessels that are being introduced into many of the world's navies. The U.S.S. *Ticonderoga* is the first of 17 advanced fleet-defense cruisers planned for the U.S. Navy. "Jump jet" aircraft carriers such as the *Invincible* have replaced large-deck carriers in the British fleet. The battleship U.S.S. *New Jersey,* which was built during World War II, was reactivated earlier this year. The Reagan Administration currently plans to build at least two more large aircraft carriers of the *Nimitz* class.

Sheffield, was one of 10 Type 42 destroyers built over the past decade for service with the British North Atlantic fleet. It and its sister ships were designed to serve as launching platforms for the new Sea Dart antiaircraft missile; their primary mission is to provide air defense for the naval forces of the North Atlantic Treaty Organization (NATO). Each Type 42 destroyer is outfitted with 20 Sea Dart missiles, a Vickers 115-millimeter automatic gun on the foredeck, two Oerlikon 20-millimeter guns aft and a helicopter armed with four Sea Skua antiship missiles on the rear deck.

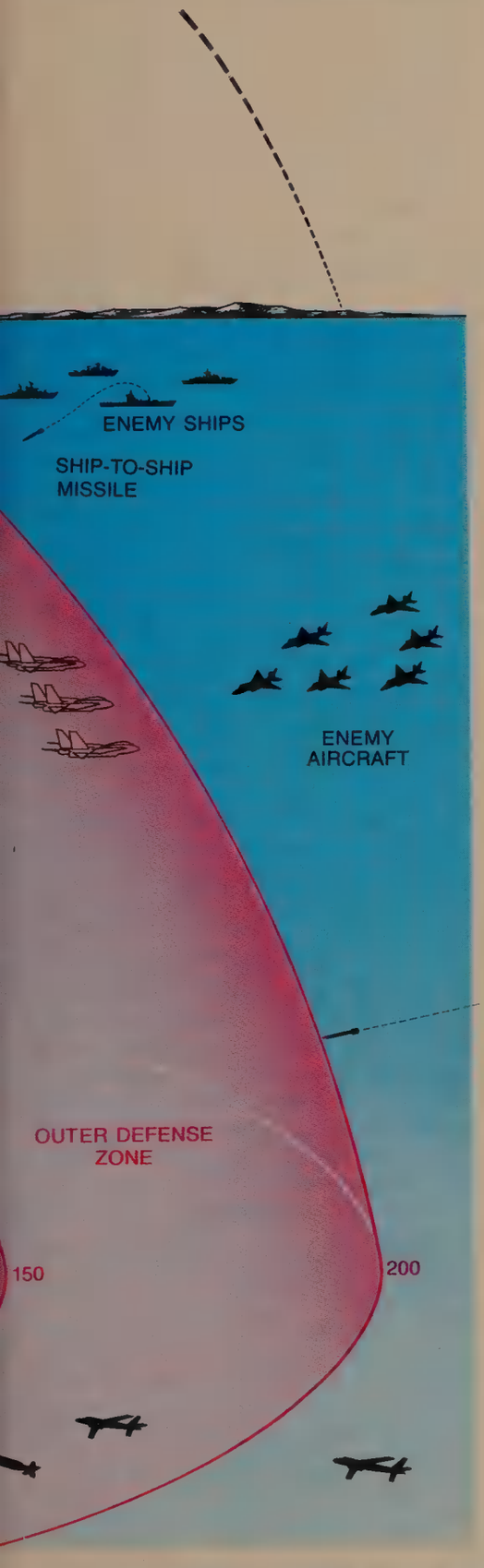
The performance of the Type 42 destroyers against attacking aircraft flying at high or medium altitudes is considered to be very good; indeed, the Sea Dart system shot down eight aircraft during the Falklands war. In the May 4 attack on the *Sheffield,* however, the attack was launched from below radar coverage, and by the time the Exocet was detected visually it was too late for the ship to respond adequately either with antimissile gunfire or with passive countermeasures.

The *Sheffield* had not been outfitted with another advanced British defensive



ZONE DEFENSE designed to protect a U.S. aircraft-carrier task force against a variety of possible attacks is depicted in this idealized scene. In the outer zone attacking aircraft, submarines and surface ships are detected and identified by surveillance satellites or by carrier-based early-warning aircraft; the attackers are intercepted by carrier-based fighters, attack aircraft and antisubmarine-warfare (ASW) aircraft and submarines. In the middle zone attacking aircraft, missiles and submarines are opposed by surface-to-air missiles and antisubmarine missiles launched by ships or helicopters. In the inner zone "point defense" systems such as rapid-fire guns and

missile, the Seawolf, which is intended specifically as a close-in, fast-reaction antimissile missile for just such sea-skimming threats. Two British frigates, H.M.S. *Broadsword* and H.M.S. *Brilliant,* were armed with the Seawolf and used it effectively in the Falklands conflict, bringing down five Argentine aircraft while they were engaged in low-level attacks. It is questionable, however, whether the *Sheffield* would have survived even if it had been armed with Seawolf missiles; the ship did not have its active search radar operating at the time of the Exocet attack in order not to interfere with communications. (Furthermore, according to a recent report, a passive radar system on board the *Sheffield* did detect the approaching missile's homing-radar beam, but the ship's computer had been programmed to identify the Exocet as a "friendly" missile and so did not sound the alarm.)



high-speed antimissile missiles serve as last-resort interceptors. The three defense zones are delineated here somewhat arbitrarily; actually there is considerable overlap between them. Ships and aircraft are not drawn to scale.

The Argentines fired only six Exocet missiles in the course of the Falklands conflict. Four of them hit targets, although perhaps not their intended ones. The first missile, which struck the *Sheffield,* was probably intended for the nearby aircraft carrier H.M.S. *Hermes;* the two ships apparently were confused on the Super Étendard's radar screen. The second Exocet, launched almost simultaneously from an accompanying Super Étendard, passed by the *Sheffield,* falling into the sea short of the *Hermes.* Three weeks later the converted merchant ship *Atlantic Conveyor,* which had delivered a load of Harrier "jump jets" to the British forces and was ferrying supplies and helicopters between the islands, was hit and destroyed by two Exocets, also launched from Super Étendards. A fifth Exocet attack five days later failed. The sixth Exocet, launched from a shore battery on June 11, struck the guided-missile destroyer H.M.S. *Glamorgan,* which was severely damaged but remained afloat.

The 67 percent success rate of the Exocet is remarkable, given the newness of the French missile system and aircraft to the Argentine forces and the much greater overall military power of the British forces. It is worth noting in this context that the Argentines had some difficulty arming the Exocets properly in time for the war. When the first missile penetrated the *Sheffield,* its warhead did not explode; because of the missile's abbreviated flight, however, it contained a substantial amount of unexpended fuel, which exploded and ignited some of the ship's fuel. It was largely a matter of chance that the missile struck in a particularly vital spot near the ship's center of operations.

The success of the Exocet in destroying a target hundreds of times its value has been widely interpreted as demonstrating the impending vulnerability of all surface warships. Nevertheless, some military analysts draw a quite different conclusion. British naval spokesmen, for example, cite the *Sheffield* incident in support of their call for an accelerated buildup of the British surface fleet. They point out that the *Sheffield* was designed for an area-defense role in coordination with large NATO fleets and the highly capable air cover of the U.S. aircraft-carrier task forces in the North Atlantic. British naval officers criticize the retrenchment in their country's defense spending over the past decade, pointing out that less than 20 years ago there were five large-deck aircraft carriers in the British navy. The last British carrier of this type, H.M.S. *Ark Royal,* was retired in 1978, leaving naval air cover dependent on the much smaller *Hermes*-class and *Invincible*-class jump-jet aircraft carriers.

The U.S. Department of the Navy takes a similar view. Secretary of the Navy John F. Lehman, Jr., for example, maintains that it should come as no surprise that one lone ship on picket duty, without adequate antimissile defenses and without adequate air cover, was struck by a homing missile of the Exocet's capabilities. He adds that "if an American ship had been the target of the Argentine attack, it would not have gotten anywhere near kill range."

On this point Lehman is probably right. The U.S. deploys its naval forces, now numbering almost 500 warships, around 13 large-deck aircraft carriers. The current plans call for increasing the number of ships to more than 600, to be organized into 15 carrier task forces and four battle groups. (Each of the latter formations is to be centered on a refitted World War II–vintage battleship.) A carrier task force is a formidable fighting group. A *Nimitz*-class, nuclear-powered aircraft carrier weighs approximately 90,000 tons, four times the displacement of either of the British carriers involved in the Falklands fighting (and more than 20 times the displacement of the *Sheffield*). The U.S.S. *Nimitz* has a flight-deck area of more than 4.5 acres and carries a total of about 100 aircraft: jet fighters, jet bombers, surveillance aircraft, antisubmarine aircraft (including helicopters), utility helicopters and rescue helicopters. At least two-thirds of the aircraft are intended to defend the fleet against attacking aircraft, surface ships and submarines. The remaining 30 or so are there to "project power" (military parlance for attacking the enemy). In addition the task force includes a number of other ships—cruisers, destroyers, submarines and supply ships—whose mission is to protect and maintain the task force for periods of up to six months at sea.

The defense of a U.S. carrier task force is divided into three major zones. In the outer zone submarines and carrier-based aircraft are the main interceptors; in the middle zone ship-launched and helicopter-launched missiles predominate; in the inner zone short-range, "point defense" systems such as rapid-fire guns and antimissile missiles are the weapons of last resort. Any attacking system—aircraft, missile, ship or submarine—must succeed in penetrating all

three zones in order to reach the command ship: the aircraft carrier.

A carrier of the *Nimitz* class typically has four E-2C early-warning aircraft equipped with long-range radar. The E-2C's fly over or near the task force at an altitude of between 25,000 and 30,000 feet, in effect extending the "vision" of the task force from about 30 nautical miles out to more than 200 miles. The E-2C is said to be capable of detecting, identifying and tracking more than 600 airborne and surface targets at a time and of directing more than 40 fighter planes from their combat-air-patrol flight paths toward individual targets.

The chief carrier-based combat aircraft in the U.S. Navy today is the F-14, a high-performance, swing-wing jet that can fly at speeds greater than Mach 2 (that is, more than twice the speed of sound). The F-14 is armed with eight precision-guided air-to-air missiles and a 20-millimeter cannon. Aircraft on combat air patrol normally "loiter" between 150 and 200 miles from the carrier, and they are reportedly able to detect bombers at a range of 170 miles or cruise missiles at 65 miles. The radar system on the F-14 that launches the Phoenix missiles can track as many as 24 targets simultaneously and can coordinate an attack on as many as six at a time at a maximum range of 60 miles. In short, the outer defense zone of a U.S. carrier task force is quite well equipped to handle multiple threats at a distance of 200 miles or more from the carrier.

The second zone of defense is based primarily on the surface-to-air missiles of the other ships in the task force. Most U.S. cruisers, destroyers and frigates now carry surface-to-air missile launchers loaded with Talos, Terrier or Tartar missiles. The newest antiaircraft missiles in the U.S. fleet, called Standard 1 and Standard 2, are just entering service. They are described as being capable of intercepting attacking aircraft at a range of between four and 85 miles and at an altitude of up to 80,000 feet. Time is of vital importance here, with a few minutes being the maximum time available to intercept a supersonic aircraft or missile in the middle defense zone.

If an incoming aircraft or missile manages to penetrate the first two defense zones, several options remain. In the final few miles of flight the attacker can be destroyed by the older antiaircraft guns still present on most ships, but in the majority of cases it will be opposed as well by modern point-defense systems. The newest point-defense missile on board U.S. ships is the NATO Sea Sparrow, a 12-foot supersonic missile with a radar-homing guidance system. The newest point-defense gun is the Phalanx, a six-ton, fully automatic Gatling gun with six rotating barrels that is reported to be capable of firing a "wall of bullets" at a rate of 50 per second. The bullets have a uranium core that makes them extremely dense and gives them high penetrating power. They are directed at the approaching target by a "closed loop" system that simultaneously tracks the projectiles as well as the target, adjusting the aim of the gun until the target is hit. If a Phalanx gun had been on board the *Sheffield,* it might have been able to stop the Exocet; the British apparently think so, since they have just installed two such guns on their newest aircraft carrier, H.M.S. *Illustrious.*

In addition to such active point-defense systems a variety of passive countermeasures are also available; they include evasive maneuvers, electronic jamming of the attacker's radar, floating infrared decoys and the dispersal of "chaff": clouds of metallic strips that confuse the radar-based guidance systems of the incoming missiles. Both chaff and decoys were used successfully by the British in the Falklands fighting.

In the continuing effort to upgrade its ship defense the U.S. Navy earlier this year commissioned its first *Ticonderoga*-class guided-missile cruiser, which is particularly well equipped to defend an aircraft-carrier task force. The Navy has proposed building 17 of these $1 billion ships, each outfitted with the new Aegis combat-control system, as many as 68 Standard 2 missiles, 20 ASROC antisubmarine missiles, two Phalanx guns and eight Harpoon antiship missiles, along with other weapons. The heart of the Aegis system is the Spy-1 phased-array radar, which is capable of sweeping the horizon in one-fourth the time of its predecessors. The system is designed to detect and track incoming objects in greater numbers and in more extreme circumstances (high and low altitudes, high and low speeds, large and small radar cross sections) with unprecedented rapidity, flexibility and firepower.

U.S. fleet defense thus appears to be highly capable. Indeed, had such a carrier task force sailed into combat in the Falklands last year, few if any Exocets would have reached their targets. Yet one's confidence in fleet defense must be tempered, particularly in view of certain operational limitations, past experience and a consideration of the potential foe.

The operational limitations of fleet defense are numerous. Assuming an attack is detected by patrolling E-2C aircraft, the maximum number of fighter aircraft a carrier can launch is 24;

NEW SURFACE-TO-AIR MISSILE, called the Standard 2, is shown being test-fired from a rail-type launcher on the deck of the guided-missile cruiser U.S.S. *Wainwright.* The Standard 1 and Standard 2 missiles are intended respectively to replace the Tartar and the Terrier missiles as the primary medium-range and extended-range antiaircraft missiles of the U.S. fleet.

each fighter in turn can be equipped with a maximum of eight air-to-air missiles. This gives a total of 192 shots at incoming aircraft and missiles, which is probably enough in limited engagements but probably insufficient in intense battles. It is for this reason among others that carrier task forces often operate in tandem, thereby doubling the number of available interceptor aircraft.

According to the Navy, the Phoenix missile on board F-14 fighters has an 80 percent kill probability. The General Accounting Office stated in a report to Congress in 1979, however, that the weapon would not be able to maintain that standard against the current generation of Russian aircraft. Another air-to-air missile in the U.S. arsenal, the heat-seeking Sidewinder, proved very reliable in the Falklands conflict, destroying 24 of the 27 Argentine aircraft it was fired at, for an 89 percent kill rate. Nevertheless, some penetration of the outer defense perimeter of a carrier task force seems inevitable. It could be accomplished through an assortment of tactics: jamming of the defensive radars, successful attacks on the E-2C's or the F-14's, the deployment of decoys or simply saturation of the defense by an overwhelming number of attacking aircraft and missiles.

The effectiveness of the middle zone of fleet defense is a function of both the speed and the capacity of the defensive combat-control system. Incoming supersonic aircraft and missiles allow only a few minutes for the defensive system to detect a target, identify it, track it, pass it on to the fire-control radar and fire the defensive missile—all before the attacker moves to within a few miles of the carrier task force.

Short of the Aegis system, the Tartar-D surface-to-air missile system is generally regarded as the most effective U.S. fleet-defense system. Yet it has serious operational limitations; for example, the rotating antennas of its search radars take four seconds to sweep the horizon, thereby setting a lower limit on the system's reaction time, an important factor in identifying and tracking a target. Moreover, the detection, identification and tracking of a target are done by human operators until the system locks onto the target, making it hard to deal with multiple targets flying at supersonic speeds. Indeed, a recent congressional study called the reaction time and firepower of the Tartar system "inadequate," referring specifically to its "considerable degradation" in the face of sophisticated electronic countermeasures. Furthermore, there is evidence of a serious lack of fire-control coordination within a task force. Several Tartar-type systems may lock onto a single incoming missile, allowing other missiles to penetrate the middle defense zone.

The Aegis control system overcomes some of these operational difficulties. The Standard 2 missile gives it greater firepower; its phased-array search radar (which points the beam electronically rather than mechanically) sweeps the horizon in only one second; it is fully automated, and it is better able to coordinate fire control within the task force. Nevertheless, the Aegis system reportedly lacks the ability to intercept sea-skimming missiles such as the Exocet. Furthermore, it is designed primarily to protect itself and any vessels behind it (with respect to the incoming missile). In operational terms, therefore, the Aegis system can handle multiple threats only within a limited sector; if attacking missiles were to arrive from several directions at once, throwing what the Navy calls "crossing targets" at the system, penetration would be virtually certain.

That leaves the inner, point-defense zone to pick up and destroy (or deflect) all the remaining aircraft or missiles. The constraints here are very tight: a missile arriving at the speed of sound (Mach 1) requires a reaction time of less than 30 seconds. The incoming missile can be fired at by either an automatic antimissile missile, such as the NATO Sea Sparrow, or an automatic rapid-fire gun, such as the Phalanx. Both have their operational limitations, however. The NATO Sea Sparrow is fired from a deck-mounted canister that contains only eight missiles and can therefore be overwhelmed. The Phalanx has an impressively high rate of fire—3,000 rounds per minute—but its magazine holds only 989 rounds, enough for less than 20 seconds of fire at the maximum rate. It too can be overwhelmed. Finally, passive countermeasures that attempt to "fool" the missile rather than destroy it can be effective in certain circumstances but not in others, as the British experience in the South Atlantic showed.

PHALANX ANTIMISSILE GUN is described as the Navy's first all-weather, fully automatic point-defense gun. The six-barrel, 20-millimeter Gatling gun fires high-velocity, uranium-core bullets at a maximum rate of 50 per second. Nevertheless, it has a number of operational limitations: its magazine holds only 989 rounds, allowing it to fire for less than 20 seconds at the maximum rate; furthermore, its barrels wear out after seven minutes of maximum firing, and the entire gun wears out after 50 minutes. The gun can therefore be easily overwhelmed.

The absolute last resort of a ship is to absorb a missile hit, as the *Sheffield* did, and to control the damage, as the *Sheffield* could not. The designers of combat ships take great care to provide adequate damage-control systems to enable a ship to survive a first hit and ideally to enable it to continue combat operations. The U.S. Navy maintains that its ships are designed to withstand several hits and still go on fighting. The hulls of most large warships are compartmentalized and are made of heavily armored steel; the superstructures, although they are often constructed of a lighter material such as aluminum, are shielded in some of the newer vessels by a layer of Kevlar, a rigid, resin-based protective armor. The steel armor of the recently reactivated battleship U.S.S. *New Jersey* is more than a foot thick in many places; it is doubtful an Exocet missile could penetrate very far into such armor.

The history of naval warfare, how-

ever, is replete with incidents in which ships have been put out of action by seemingly minor hostile actions or shipboard accidents. The German admiral Alfred von Tirpitz once remarked that a warship has three goals: first, to stay afloat; second, to stay afloat; third, to stay afloat. Large modern warships may well be able to do that after being hit by a smart weapon or even by several weapons, but to carry on normal operations may be another matter. Most naval officers readily admit that in modern warfare a large ship will stay out of the line of fire until the enemy is "neutralized," most likely by aircraft.

The vulnerability of the U.S. Navy is most apparent when one considers the maximum potential threat it faces: not a half-dozen Exocet missiles fired by a few Super Étendard aircraft but the Russian fleet of more than 700 surface warships (mostly limited-range frigates, corvettes and coastal patrol vessels), almost 300 submarines (mostly diesel-powered) and more than 1,500 antiship cruise missiles. The missiles, which are based on surface ships, submarines, aircraft and shore batteries, have ranges of up to 350 miles and speeds as high as Mach 4. The Phalanx gun, which is designed to stop a missile in the final kilometer of its flight, would have slightly more than one second to hit an incoming Mach 4 weapon. Although many of these missiles would not be encountered outside the coastal zones of the U.S.S.R., some of the longer-range Russian aircraft, including the Backfire bomber, are capable of projecting force at greater distances.

Another critical factor in any assessment of the vulnerability of the U.S. surface fleet is the existence of nuclear weapons. Most of the antiship missiles in the Russian arsenal can carry either a nuclear warhead or a conventional (non-nuclear) one. No ship, no matter how large or well armored, can survive a nuclear explosion nearby.

In short, the advent of smart missiles, with or without nuclear warheads, is bound to place a premium in naval warfare on stealth, deception, mobility and dispersion rather than on size, armor plate or coordinated defensive firepower. To be sure, the technological advantage will continue to alternate somewhat between offense and defense, as it has in the past. Antiship missiles will continue to become more accurate, more maneuverable, longer in range, faster and stealthier, and so will defensive missiles. At the same time ships will continue to evolve into more effective platforms for improved radar and fire-control systems, rapid-fire guns, antimissile missiles and passive countermeasures. In the long run, however, surface warships must become increasingly expendable.

This conclusion leads one to question the current naval strategy of the U.S. The declared policy is one of mari-

ULTIMATE ANTISHIP WEAPON currently available in the arsenals of both the U.S. and the U.S.S.R. for possible use against the other side's naval forces is a long-range precision-guided nuclear warhead. The effects of such a hypothetical attack are suggested by this photograph of the Baker test explosion conducted by the U.S. at Bikini atoll in 1946. Some 70 unmanned ships of various types were anchored around the site of the shallow underwater burst, which had an explosive yield of about 20 kilotons (small by today's standards). The

time supremacy based on large-deck aircraft carriers. Secretary of the Navy Lehman explains: "Carriers are the only way to get superiority over seven-tenths of the earth's surface." He and Secretary of Defense Caspar W. Weinberger argue that the Falklands experience only strengthens the case for additional carriers and that it did not teach any "brand-new lessons."

The lessons may not be new, but they are cautionary nonetheless. Capital warships are increasingly at risk, regardless of their armament, to comparatively inexpensive smart missiles and torpedoes. For a carrier task force to be effective against a capable opponent such as the U.S.S.R. one must assume that (1) the fleet will remain hidden, (2) virtually all the attacking missiles and aircraft will be defeated and (3) nuclear warheads will not be employed. Given the increasing effectiveness of satellite and aerial reconnaissance, the permeability of fleet defense and the strong reliance of the Russians on nuclear weapons in their military planning, none of these assumptions appears plausible.

It may be, therefore, that the U.S. Navy considers its fleets expendable. Retired Admiral Hyman G. Rickover, the "father" of the nuclear navy, testified before a congressional committee last year that a modern aircraft carrier would last "about two days" in a major war. The Joint Chiefs of Staff project the need for 25 such carriers for "minimal risk"; that number or more might be needed just to keep pace with attrition in a major war. Yet a new, 90,000-ton aircraft carrier that costs more than $3.5 billion to build (in 1982 dollars) and another $15 billion or so to outfit and maintain—and that has a crew of more than 6,000—is not an obviously expendable item.

In meeting their stated objectives of "controlling the seas" and "projecting power" aircraft-carrier task forces are also intended for foreign intervention. As sophisticated antiship technology proliferates into the Third World, however, even small nations will represent a serious threat to large surface vessels. To cite just one example, Libya is reported to have 13 kinds of smart missiles in its arsenal.

Perhaps the final lesson of the Falklands war is that when it happens again, it will be even costlier. The days of gunboat diplomacy are over, and yet some of the major powers fail to recognize the fact. Military custom dies slowly. "In harm's way," a traditional American naval motto, will eventually give way to the recognition of the growing dominance of the naval conflict by land-based aircraft and missiles, by small, mobile, fast (and expendable) ships armed with long-range precision-guided weapons and by submarines. Aircraft carriers, reactivated battleships and other large, expensive surface vessels will retain one safe role: to show the flag in peaceful waters.

ships suffered damage from a number of causes, including underwater and atmospheric shock waves as well as surface water waves. For example, the aircraft carrier U.S.S. *Saratoga,* which was moored in the lagoon about 400 yards from "surface zero," had the central part of its island structure folded down onto the deck by a 90-foot wave generated by the explosion. The radiation effects on the ships' crews in the event of such an attack would of course extend far beyond the immediate destructive effects on the ships themselves.

Vibrations of the Atomic Nucleus

The nucleus can quiver, ring or even "breathe"; the coordinated motion of the nuclear particles reveals much about the forces between them. Six modes of vibration have been detected so far

by George F. Bertsch

Vibrating systems, from the swinging pendulum to the oscillating electromagnetic field of a light wave, have long had a privileged place in the physical sciences. The vibrational modes and frequencies observed in a system can reveal much about the nature of the forces acting within it. An analysis of vibrational motions was notably important in the understanding of atomic structure in the early years of the 20th century. More recently a richly varied spectrum of vibrations on a far smaller scale has been discovered in the nucleus of the atom. The study of the nuclear vibrations is proving to be a major source of information on the structure of the nucleus and on the forces that hold it together.

The protons and neutrons that make up a nucleus are collectively called nucleons. Quantum mechanics describes the arrangement and the motion of the nucleons by means of a wave function. For the stable nuclei found in nature the wave function does not change with the passage of time. The nucleus can be set in motion by external forces, however, leaving it internally excited. In the simplest excitations, called giant vibrations or giant resonances, all the nucleons oscillate coherently and the motion follows a simple pattern. The difference between coherent and incoherent motion in an excited nucleus is roughly analogous to the difference between the coherent motion of the liquid in a cup of tea that has been bumped and the random thermal motion of the molecules in the hot tea. In a coherent vibration there is a pattern: like liquid sloshing in a cup, the nucleons in a vibrating nucleus pass cyclically from one distribution to another.

The motion of the individual particles in a vibrating body can be coordinated in various ways, giving rise to the distinctive patterns of motion called vibrational modes. The molecules in the cup of tea can oscillate from the side of the cup to the center and back again or from one side of the cup to the other, and the nucleons can also oscillate in several patterns.

So far six giant nuclear vibrational modes have been observed. Most of them have been detected experimentally only in the past few years. The giant vibrational modes fall into two classes. The first class consists of the modes the nucleus shares with other spherical bodies, such as a drop of water or the earth itself. The spherical vibrational modes of the nucleus that have been observed are, in the order of their discovery, the giant dipole, the giant quadrupole, the giant monopole and the giant octupole. The terms were borrowed from the names for electric fields of a given spatial complexity. For example, the protons in a nucleus vibrating in the quadrupole mode give rise to an oscillating electric field that resembles the field generated by four poles, or point charges.

The other class of nuclear vibrations involves the spin orientation of the nucleons. Protons and neutrons, like electrons and many other particles, spin on an internal axis. In the unexcited nucleus the spin axis is fixed with respect to the wave function of the nucleon and the spin orientation of the other nucleons. During a spin vibration some of the nucleon spins are tipped slightly and the spin axes begin to precess, that is, they describe circles around their original orientation. The precession of neutrons with respect to protons can be coordinated in different ways, giving rise to different spin-vibrational modes. For example, the neutron spins can be tipped in the same direction as the proton spins or in the opposite direction. As a result of the coordinated precession of the nucleons the nucleus can acquire a net spin or a net magnetic moment that oscillates at the precession frequency. So far two vibrations of this kind have been observed: they are called the giant Gamow-Teller resonance and the giant magnetic-dipole resonance.

An understanding of nuclear vibrations requires knowledge of the forces acting between nucleons and the mechanical laws governing the motion of the nucleons under the influence of these forces. Given the complexity of the problem in the many-body system of the nucleus, physicists have long resorted to simple models. The nucleons are viewed as moving under the influence of a single generalized force whose effects approximate those of the interactions of the many nucleons. The vibrational modes of the nucleus provide an experimental check on the validity of the models; a theoretical justification requires the more fundamental quantum theory.

A body vibrating in a given mode oscillates at a specific frequency called the resonance frequency. The resonance frequency is determined in part by the internal forces opposing the motion of the oscillating particles. A drop of water and a solid sphere vibrating in the same spherical mode have different resonance frequencies. The solid strongly resists any distortion and has a high vibrational frequency. The liquid drop, on the other hand, returns to a spherical shape only because of forces associated with surface tension. These forces are quite weak and lead to a low vibrational frequency.

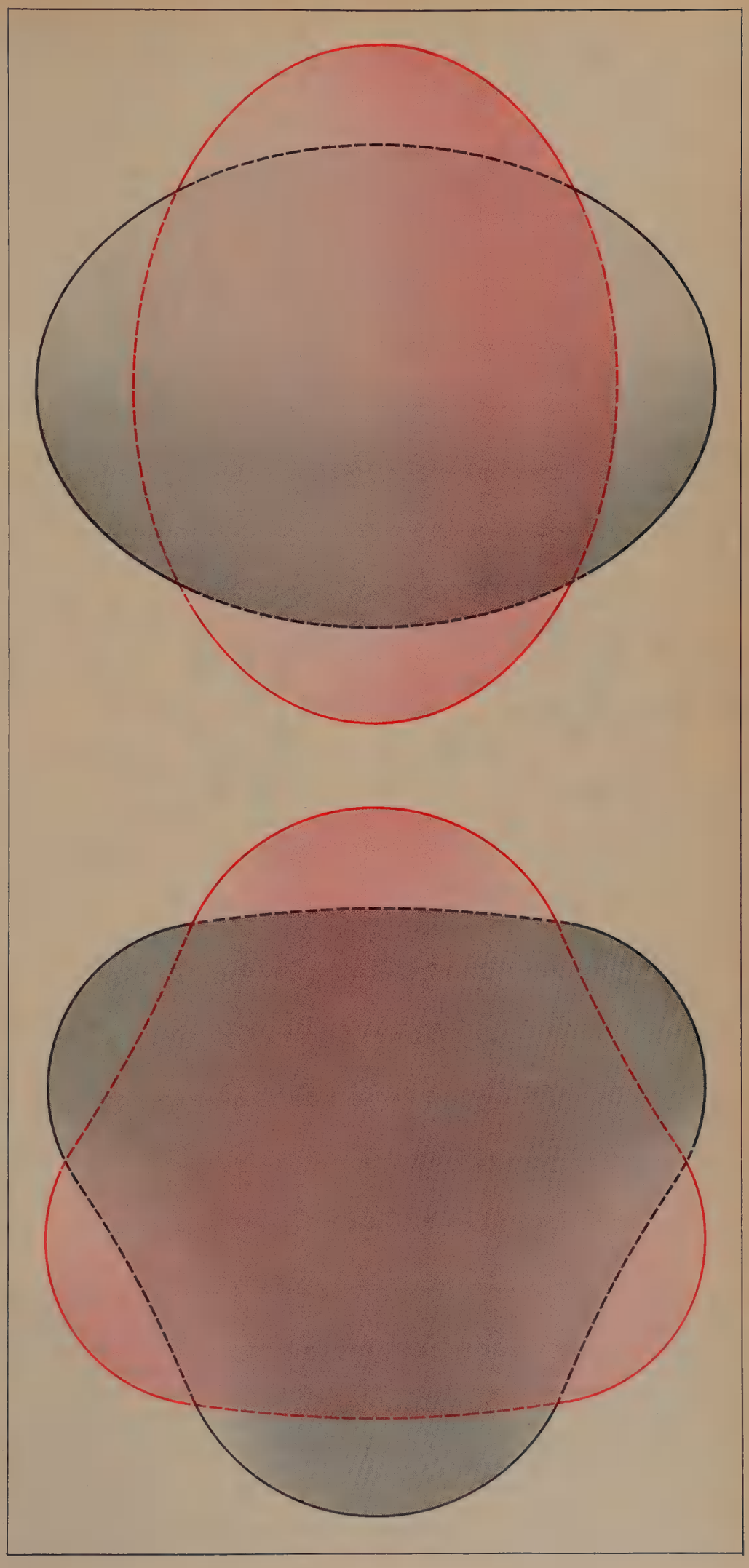
One model of the nucleus, the liquid-drop model, likens the forces acting between the nucleons to those acting between the molecules of a low-viscosity liquid. Another model, based on the shell model of the nucleus, likens the forces to those between the particles in an elastic solid. The resonant frequencies of nuclear vibrational modes were calculated from the liquid-drop model and from other simple models long before the vibrations could be excited in the laboratory and their frequencies measured. No single model accurately predicts the resonant frequencies of all modes of vibration. In some circumstances the nucleus acts like a liquid and in others like an elastic solid. In general its response is rather like that of a class of non-Newtonian fluids, of which Silly Putty is the most familiar example.

These substances respond elastically to sudden forces but flow as liquids over longer periods of time.

The most accurate and sophisticated description of nuclear vibrations is given by the time-dependent mean-field theory, which is based on quantum mechanics. In the mean-field theory the forces acting on the nuclear particles are calculated from the quantum-mechanical motion of the particles themselves, whereas the simpler models are based on specific assumptions about the relation between the forces and the particle motion. At first sight these approaches seem quite divergent, but the motion in the vibrations is simple, and in many cases it can be shown through the quantum theory that the assumptions made in the simple models are justified.

In general, nuclear vibrations are excited by bombarding nuclei with high-energy photons (quanta of electromagnetic radiation) or other particles. The vibrations are detected by observing how the projectile is absorbed or diffracted by the nuclei. In the case of the first vibration to be reported, the giant dipole vibration, both the excitation and the detection of the vibration proved to be relatively straightforward. For the other modes it has proved quite difficult to probe the nucleus in just the way needed to excite a particular vibration. Indeed, nearly 30 years passed between the discovery of the giant dipole vibration and the discovery of another giant vibrational mode.

If a nuclear vibration is to be excited, the first and simplest condition to be met is that the energy imparted to the nucleus must be equal to the energy associated with the vibration. The nucleus vibrates at extremely high frequency; the energy of the vibration is equal to the frequency multiplied by Planck's constant, and so the energy is also comparatively high. A typical vibrational frequency is 5×10^{21} hertz, which corresponds to a vibrational energy of 20 million electron volts (MeV). The energy of the photon or other particle that



VIBRATIONS OF NUCLEAR SHAPE, in which the nucleus is deformed from a spherical to an ellipsoidal or pear shape, are akin to the vibrations of a macroscopic body such as as a drop of water. In both cases the vibration is a coherent oscillation of the particles of which the body is composed and hence gives rise to a pattern of motion of the body as a whole. The vibrations have had an important role in testing models of nuclear structure. Various models differ in their predictions of the frequency and the other characteristics of a given mode of vibration in the nucleus of a given atom. It seems that the nucleus sometimes acts like an elastic solid, sometimes like a liquid and sometimes like Silly Putty.

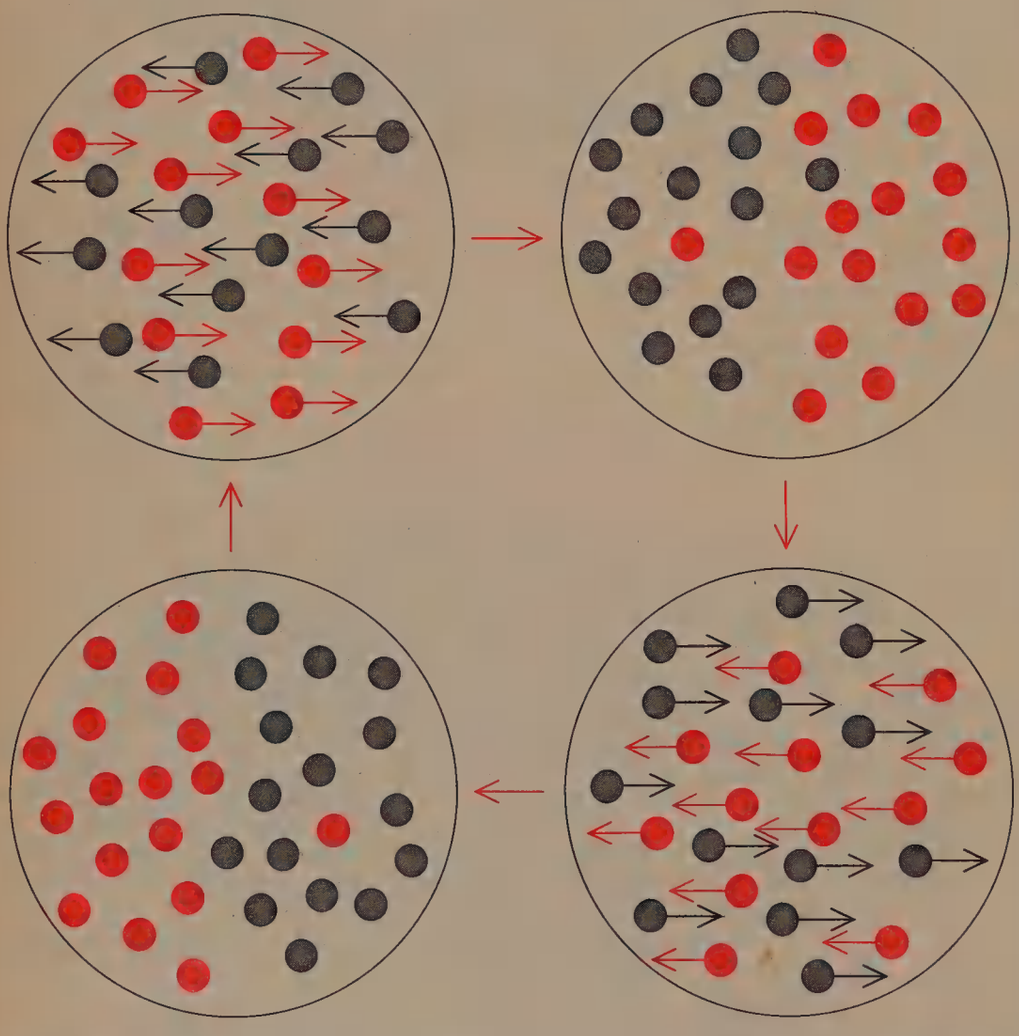
excites a nuclear vibration must be at least as high as the vibrational energy.

In the case of the giant dipole resonance the energy requirement was the only condition that had to be met. Physicists were able to excite the giant dipole vibration simply by bombarding nuclei with photons having an energy equal to the vibrational energy of the mode. The discovery of the giant dipole depended only on the availability of a source of monoenergetic high-energy photons, or gamma rays. The detection of the vibration followed closely on the development of such sources (bremsstrahlung, or braking radiation, from electron accelerators) in the mid-1940's.

It is not difficult to understand how gamma rays excite the dipole vibration. A photon carries with it an oscillating electric field. Although the wavelength of a gamma ray is smaller than that of other forms of electromagnetic radiation, such as visible light, it is large with respect to the diameter of a nucleus. As a result the electric field associated with a passing gamma ray is nearly uniform across the nucleus. The field exerts a force on the positively charged protons, moving them away from the neutrons. The neutrons themselves are electrically neutral, and so the field has no direct influence on them. Because the center of mass of the nucleus remains at rest, however, the neutrons move in the opposite direction. The restoring force of the vibration is the attractive force between protons and neutrons, namely the strong nuclear force responsible for binding the particles together. The strong force is independent of electric charge.

The giant dipole was not only relatively easy to excite but also relatively easy to detect. The photons that excite the vibration are simply absorbed in the nucleus. This phenomenon, which is a form of resonance, arises in any vibrating system excited by an external source. A person singing in a shower provides a familiar example of resonance. When the frequency of the voice matches a natural vibrational frequency of the air mass in the shower, the amplitude of the vibration becomes very large. The sound intensifies; the purity of the frequency is perceived in a resonant tone. In such situations the large amplitude leads to an increased absorption of energy. Similarly, when the frequency of the oscillating electric field associated with the gamma rays matches the resonance frequency of the dipole mode of the nucleus, the gamma rays are absorbed.

The tendency of the nucleus to absorb incident particles is expressed quantitatively as an effective cross section, measured in units of area. The dipole resonance is seen in a sharp increase in the absorption rate of the target at a particular photon energy. This is expressed as an increase in the effective absorption cross section of the nucleus. It is as if the nucleus suddenly grew larger and therefore intercepted more photons, although what actually changes is not the nuclear size but the absorbance. Because the response of a body to an electromagnetic force depends on the electric charge of the body, it is possible to determine from the effective cross section of the vibrating nucleus that all the protons in the nucleus are participating in the vibration; in short, the dipole vibration is in fact a giant vibration.



GIANT DIPOLE VIBRATION, identified in the late 1940's, is excited by bombarding the nuclei in a target with high-energy photons. The protons in a nucleus (*color*) are accelerated in one direction by the electric field associated with a passing photon. The neutrons (*gray*) are unaffected by the field, but they move in the direction opposite to that of the protons so that the center of mass of the nucleus remains stationary and momentum is conserved. The restoring force, which ultimately reverses the motions of the protons and neutrons, is the strong nuclear force responsible for binding them together. The frequency of the vibration can be calculated from a simple model in which only the inertia of the nucleons, the restoring force and the laws of classical mechanics are considered. The inertia of the system is proportional to the number of nucleons and thus to the volume of the nucleus. The restoring force is proportional to both the volume of the nucleus and the displacement of the nucleons. The dependence of the variables on nuclear volume suggests that the frequency of the vibration should vary inversely with the radius of the nucleus; the experimental data on large nuclei bear out this conclusion. Here the amplitude of the vibration has been greatly exaggerated for the purpose of clarity.

The next vibration to be reported was the giant quadrupole vibration. The quadrupole, unlike the dipole, is a shape vibration, in which the shape of the nucleus as well as the distribution of the nucleons changes. A nucleus vibrating in the quadrupole mode is distorted from a spherical shape to an ellipsoidal shape and moves back through a spherical shape to an ellipsoidal shape of another orientation. It is intuitively clear that the vibration could be induced by pushing in on the nuclear surface along one axis and pulling out along a perpendicular axis.

In order to excite the shape vibrations it is necessary not only to impart a given energy to the nucleus but also to distribute the energy in such a way that the nucleons are set in motion in different directions. Even though the vibrational energy of the quadrupole mode is from 10 to 20 MeV, within the energy range of gamma rays, they do not excite the quadrupole mode because they cannot meet the second condition. Gamma rays interact with the nucleus through the electromagnetic force; they can accel-

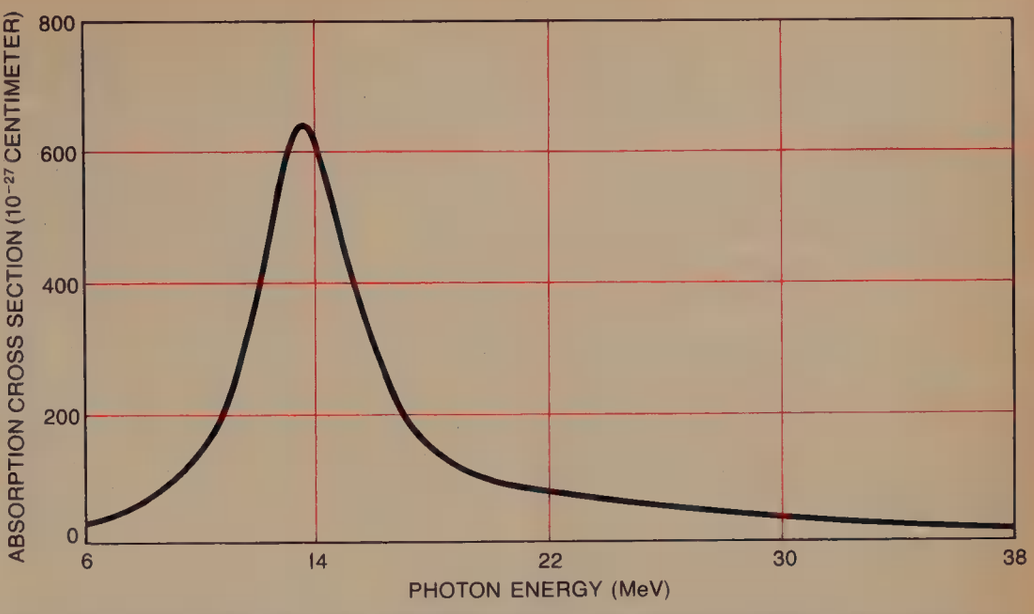
erate only the protons and, given the wavelength of the gamma rays, accelerate them only in one direction.

One answer to the problem is to excite the vibration by the inelastic scattering of a particle from the nucleus rather than by a complete absorption process. In classical physics a particle that strikes an extended object and bounces off can cause localized motion and leave the object vibrating. A similar phenomenon can take place in a quantum-mechanical system. The particles that can serve as projectiles for inducing nuclear excitations include electrons, protons and more massive nuclear particles such as the alpha particle (the nucleus of the helium-4 atom).

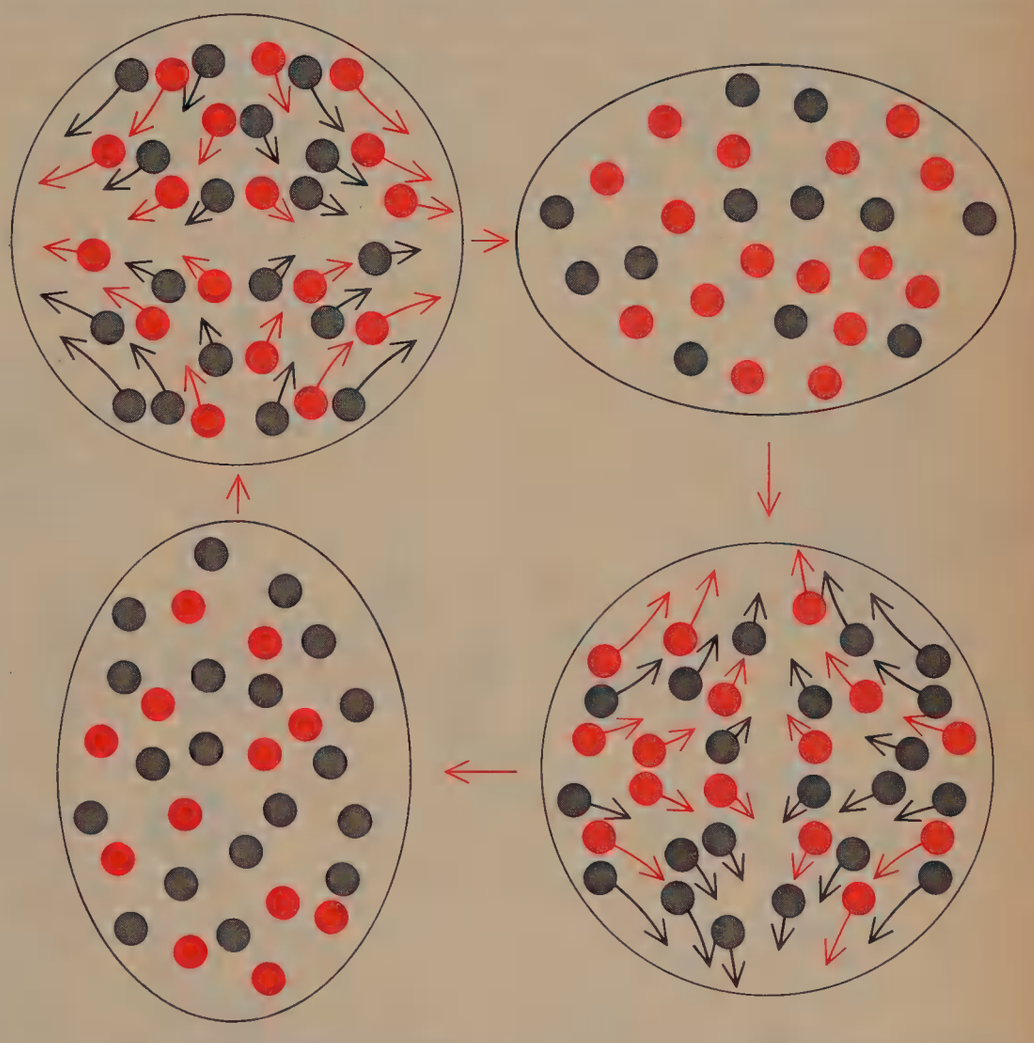
There are several advantages to choosing nuclear projectiles such as protons or alpha particles to excite the nucleus. The principal interaction in such events is mediated by the strong nuclear force, which is charge symmetric, that is, the same for protons and neutrons. The charge symmetry makes it easier to excite vibrations such as the quadrupole in which the protons and neutrons move together. Vibrations in which protons move opposite to neutrons, such as the dipole, are not excited at all by charge-symmetric forces. Thus to observe the quadrupole vibration without interference from the dipole, it is better to employ strongly interacting projectiles such as alpha particles rather than electrons, which interact with the electric field and can give rise to both kinds of vibration.

Another advantage of nuclear projectiles is that the different shape modes can be selectively excited. A fast proton going by a nucleus transmits a fleeting impulse to the nearby target nucleons, causing them to move in the wake of the projectile. The range of the strong nuclear force is small compared with the size of a nucleus, raising the question of how a coherent vibration of the entire nucleus can be set in motion. To understand how particular vibrations can be selected by the scattering process requires one of the basic notions of quantum mechanics, the wave aspect of particles. A proton going by a nucleus can be thought of as a wave enveloping the nucleus; the wave is subject to the laws of diffraction like any other wave.

Depending on how the wave interacts with the nucleus, it is diffracted with a characteristic pattern. For example, a light wave passing over a black sphere is diffracted into the shadow region, forming a pattern of rings in the center of the shadow. In the case of nuclear projectiles and targets the diffraction pattern depends on several factors. In order to produce a clear diffraction pattern at all, the wave must be absorbed in the interior of the nucleus. In physical terms, when the projectile penetrates the nu-



PEAK IN THE ABSORPTION CROSS SECTION of lead nuclei is evidence that they are vibrating in the dipole mode. Cross section is a measure of the fraction of the photons absorbed by nuclei. It increases dramatically when the photon frequency matches a vibrational frequency of the nuclei, a phenomenon called resonant absorption. The resonance in this nucleus is at a photon energy of 14 million electron volts (MeV), equivalent to a frequency of 3×10^{21} hertz.



GIANT QUADRUPOLE VIBRATION can be excited by bombarding nuclei with projectiles that interact with protons and neutrons equally by means of the strong force. The quadrupole vibration therefore differs from the dipole in that the protons and neutrons move together rather than in opposite directions. As in the case of the dipole mode, the vibrational frequency can be calculated from the inertia of the system, the restoring force and the laws of classical mechanics. Again the inertia is proportional to the volume of the nucleus, but it is not immediately clear how to describe the restoring force. If the restoring force is assumed to be analogous to surface tension, as in a vibrating drop of water, the force is proportional to the surface area of the nucleus. If the restoring force is assumed to be analogous to stress energy, or resistance to deformation, as in a vibrating elastic solid, the force is proportional to the volume of the nucleus. Experimental data indicate that the frequency of the giant quadrupole vibration varies inversely with the radius of the nucleus, which favors the model based on an elastic solid.

cleus, it must interact so strongly that it does not emerge intact. The position of the rings in the diffraction pattern is determined by the size of the nucleus, the wavelength of the projectile and the type of vibration being excited.

The interaction of the wave with the nucleons on the nuclear surface alters the form of the wave. After the interaction each small area of the nuclear surface acts as the origin of a spreading wavelet. The phase of the wavelet depends on the motion of the originating surface. Wavelets originating from outward-moving surface areas have the same phase, or sign, whereas wavelets from inward-moving surface areas have the opposite phase. As the wavelets propagate outward from the nuclear surface they overlap and interfere. Whether the interference is constructive or destructive at any point depends on the phase of the various wavelets at that point.

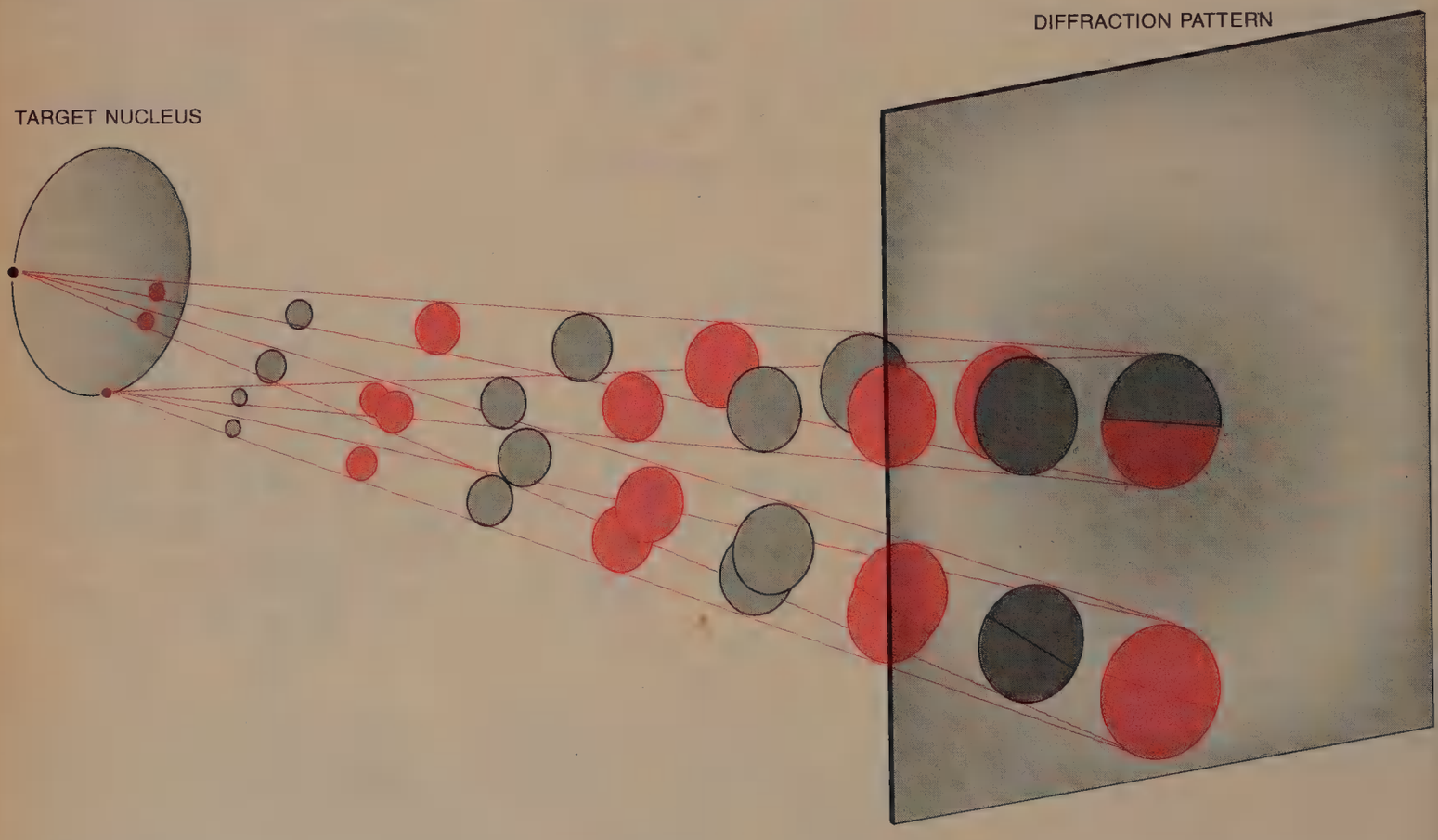
The nature of the diffraction pattern can be understood by considering a plane "downstream" from the target nucleus and perpendicular to the beam axis. If the phase of the wave were not altered by its passage near the nucleus, the phase of the wavelets reaching a given position on the plane would be determined entirely by the distance from the point of diffraction to that position. Of particular importance, wavelets diffracted to the center of the plane along the beam axis in the shadow of the nucleus would all travel the same distance and would arrive in phase. The resulting constructive interference would create a bright spot (a region of large wave amplitude) in the center of the plane.

When the wave stimulates a quadrupole vibration, however, the phases do not remain unaltered. On the contrary, the wavelets originating from the elongating sides of the nucleus are initially 180 degrees out of phase with the wavelets originating from the contracting sides. As a result the interference in the center of the plane is destructive. There are other points in the plane, however, where the distance to an elongating region and the distance to a contracting region differ by exactly half a wavelength. The wavelets arrive at these points in phase, interfering constructively. Because the axes of the inward- and outward-deforming regions of the nucleus can have any orientation, the strongest diffraction is directed into a ring centered on the axis of the beam.

The amplitude of the wave function of a particle at any point in space determines the probability that the particle will be found at that point. Consequently the projectile particles are most likely to be scattered into regions of constructive interference, where the amplitude of the wave function is largest.

Like absorption, inelastic scattering can be measured in terms of a cross-sectional area. Even in the absence of a vibrational resonance a certain fraction of the projectiles bombarding a target are scattered inelastically, and the fraction scattered in any one direction varies smoothly with the angle between the beam axis and the scattering direction. The quadrupole vibration shows up as a sharp increase in the number of particles scattered inelastically in particular directions.

For a quantitative prediction of the cross section as a function of scattering angle the interaction of the wave with the nucleus is described by a phenomenological optical model. The wave equation is then solved numerically to find the precise cross section. The cross section for the quadrupole mode reaches its maximum value at an angle that depends on the size of the nucleus as well as on the wavelength of the projectile. For example, in the case of the lead 208 nucleus bombarded by 100-MeV al-



ANNULAR DIFFRACTION PATTERN formed when projectiles are scattered by nuclei in the target is associated with the quadrupole vibration. The projectile particle can be described quantum-mechanically as a plane wave perturbed by its interaction with the surface of the nucleus. The interaction of the wave with the nucleons is attractive along one axis, so that the nuclear surface is pulled outward, and repulsive along the perpendicular axis, so that the surface is pushed inward. The single plane wave is diffracted into many small circular waves that propagate outward and interfere with one another. The waves originating from the inward-deforming regions of the nucleus are 180 degrees out of phase with those from the outward-deforming regions. (Here the phase differences are represented by color differences.) The pattern created by the interference of the waves is determined by this phase difference. Waves arrive at any point along the beam axis out of phase and cancel. There are points off the axis, however, where the waves arrive in phase and interfere constructively.

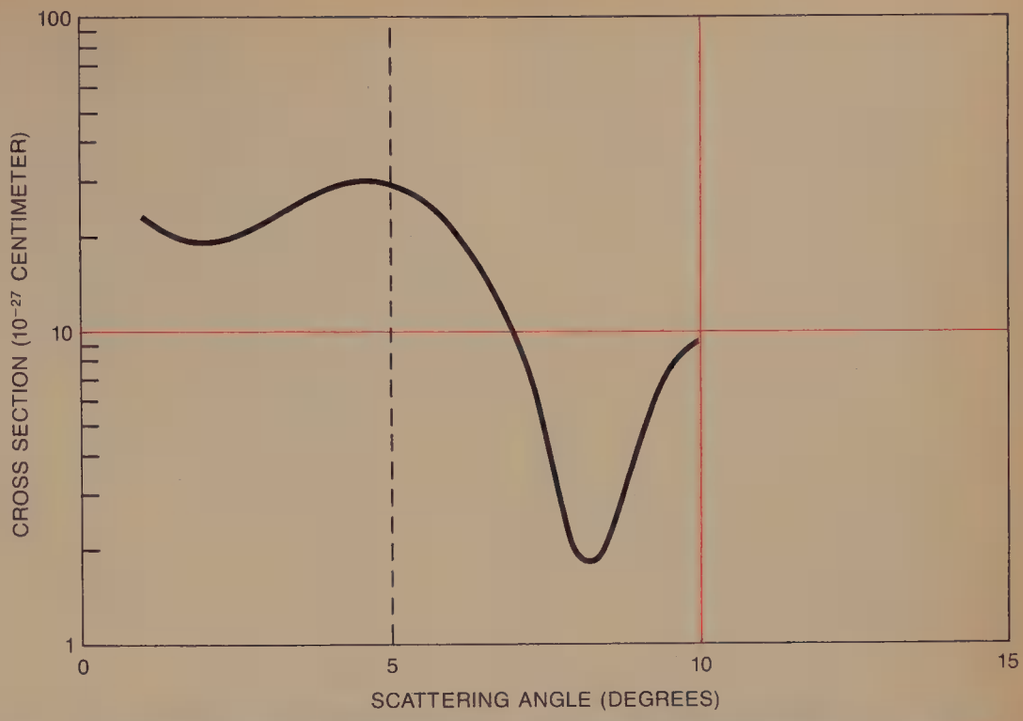
pha particles, the main diffraction peak is at five degrees from the beam axis.

The third nuclear vibrational mode to be observed, the giant monopole vibration, is excited and detected in much the same way as the giant quadrupole. The monopole vibration is a "breathing" mode: the nucleons move inward and outward from the center of the nucleus in phase with one another, so that the nucleus expands and contracts. Intuitively it would seem that the best way to induce this motion would be to push uniformly on the surface of the sphere or, equivalently, to pull outward in a radial direction.

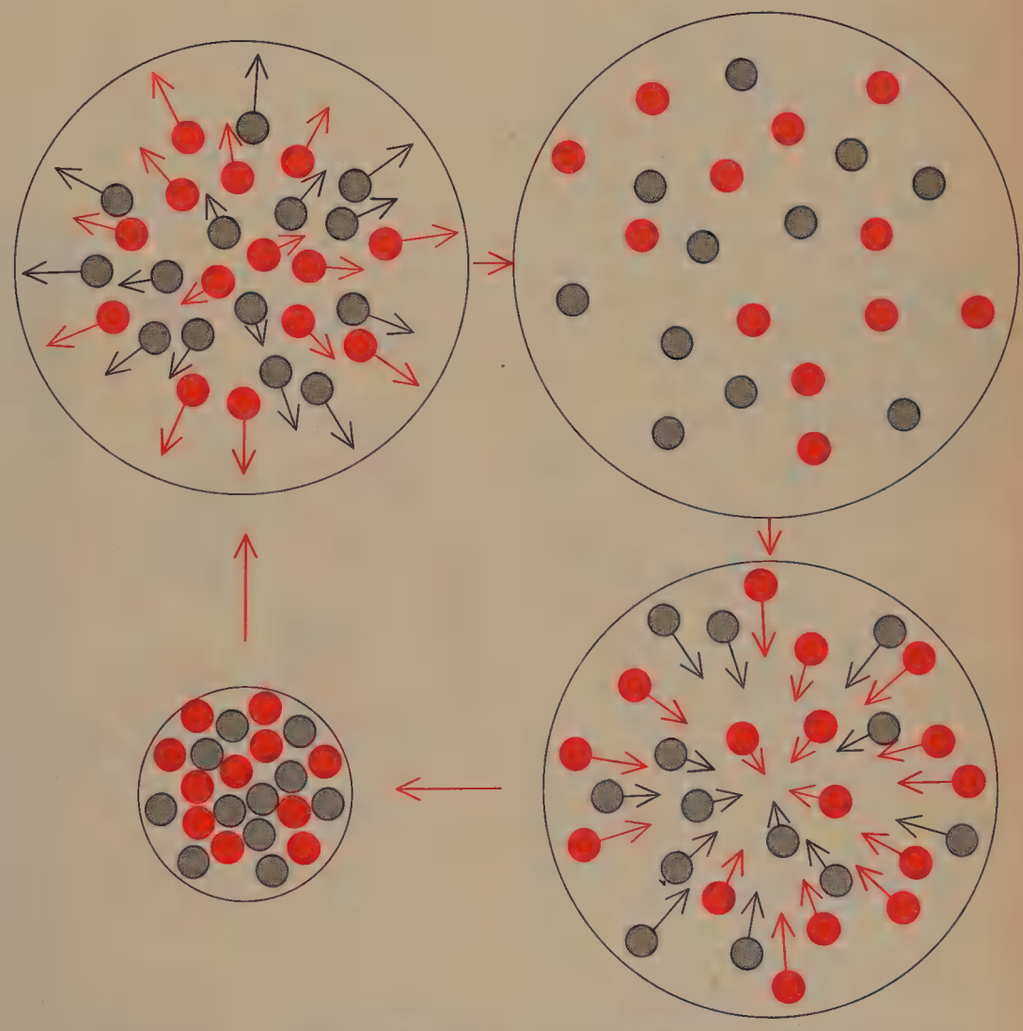
Such perfect conditions for exciting the monopole vibration cannot be achieved with inelastic scattering. If the projectile hits the nucleus, it loses most of its energy and leaves the nucleus in a very highly excited state. Only when the projectile grazes the surface of the nucleus is the interaction gentle enough to excite the simple vibratory motion. The part of the surface the projectile wave acts on is a ring-shaped region circling the beam axis. As in the quadrupole excitation, the interaction gives rise to many wavelets originating from each small area of the nuclear surface. The wavelets interfere as they spread out from the surface of the sphere. Because the monopole excitation is completely symmetric, the wavelets start out from the curved surface with the same phase. Hence along the beam axis, where the distance to all areas on the periphery of the sphere is equal, the wavelets arrive in phase and interfere constructively.

The diffraction pattern formed by the particles exciting the monopole vibration is a prominent spot centered on the beam axis and surrounded by faint rings. In other words, the monopole vibration is characterized by a maximum in the scattering cross section at zero degrees from the beam axis. The pattern is essentially the same as the one formed by particles diffracted by a nonvibrating nucleus. Indeed, it was largely for this reason the monopole vibration was harder to detect than the quadrupole vibration; it proved difficult to separate the particles that are inelastically scattered along the central axis of the beam from the particles that pass through the target without interacting with the nuclei. The only way to discriminate between the inelastically scattered particles and the beam particles is by measuring the energy dependence of the inelastic scattering.

At this point it is possible to return to the question of why it took so long to obtain experimental evidence of quadrupole and monopole vibrations. Rough predictions of the vibrational frequencies of these modes had been



PERCENTAGE OF PARTICLES scattered at ■ given angle to the beam axis increases sharply when the quadrupole vibration is excited. The scattering cross section varies with angle in a way that reproduces the diffraction pattern associated with the quadrupole mode. The scattered particles have ■ specific energy, namely the energy of the impinging beam minus the vibrational energy of the quadrupole mode. In this case the projectiles are alpha particles with an energy of 96 MeV striking a target consisting of the isotope samarium 144. The quadrupole mode is preferentially excited with ■ scattering angle of five degrees to the beam axis.



GIANT MONOPOLE VIBRATION is also called the breathing mode: in it the nucleus expands and contracts. In the model of the giant monopole vibration based ■■ classical mechanics the restoring force is the resistance of nuclear matter to compression. The observed frequency of the monopole vibration gives ■ value for the compressibility coefficient of nuclear matter.

made from theoretical models of the nucleus years before the vibrations were seen in the laboratory. Thus experimental physicists had known for some time approximately where to look and what they could expect to see. What they lacked were instruments of sufficient power and sensitivity, in combination with accelerators providing projectiles of sufficient energy.

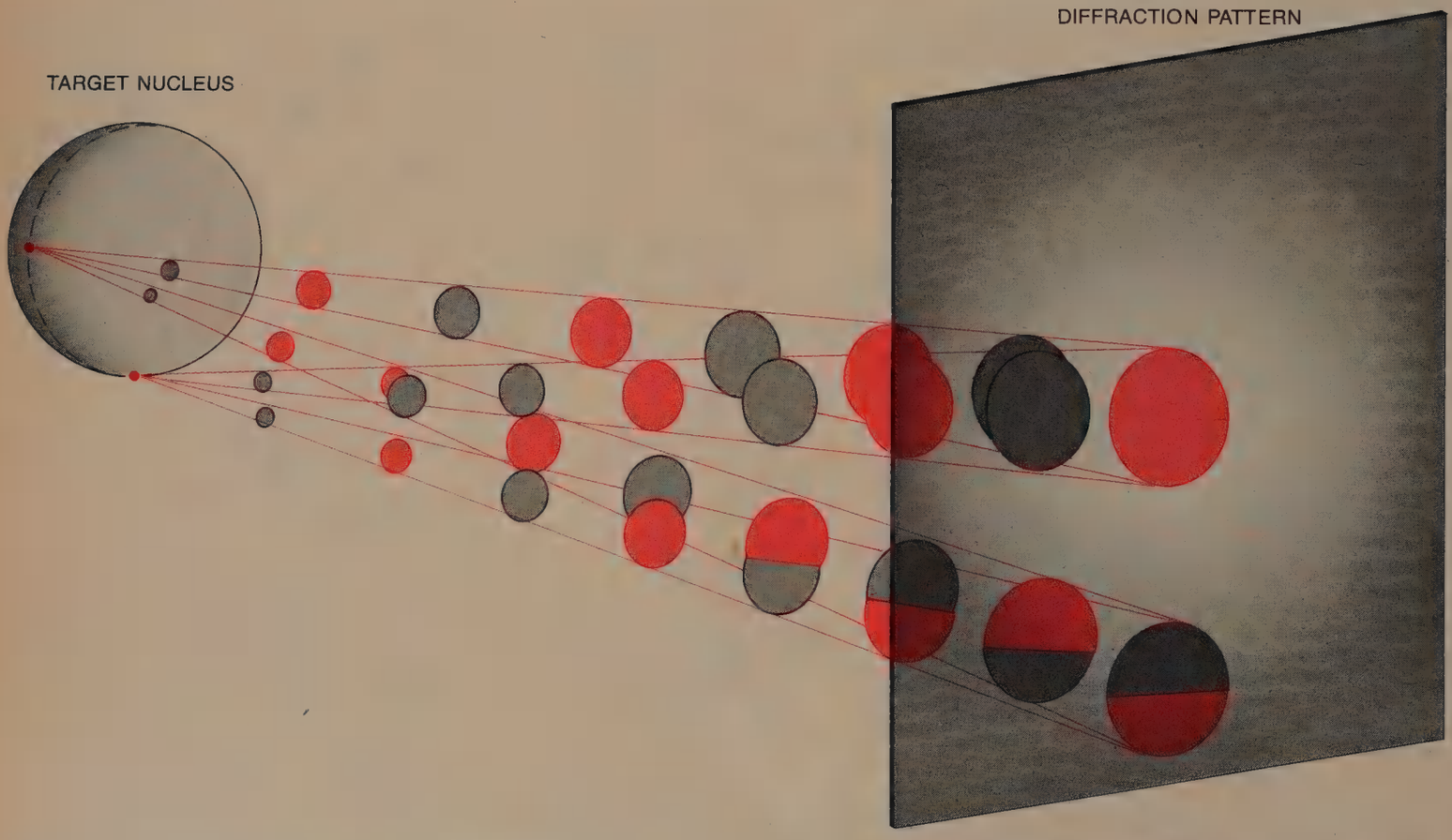
Even though the vibrational energy of the quadrupole and monopole modes lies in the range from 10 to 25 MeV, the projectile energy must be several times higher. The velocity of the projectile is an important consideration in judging its suitability as a vibrational probe. The velocity of a particle of a given mass depends on its energy. High-energy projectiles are needed to excite giant vibrations because lower-energy (and hence lower-velocity) projectiles interact with the target for a relatively long time, allowing more complex excitation processes to take place. A slower projectile might induce a vibration composed of several fundamental modes, or it might induce more complicated motions in the nucleus by exchanging nucleons with the target. These processes do not give rise to a diffraction pattern but contribute to a background cross section that must be subtracted from the data before the vibrational modes can be analyzed.

High-energy projectiles are needed for another reason. Simple diffraction patterns are obtained only if the wavelength, and therefore the energy, of the exciting projectile does not change significantly during the excitation process. A projectile that excites a nuclear vibration gives up an amount of energy equal to the energy of the vibrational mode; the higher the initial energy of the projectile, the less significant the resulting change in wavelength. On the other hand, from the experimental point of view the small change in energy makes it difficult to distinguish the diffracted particles from the much larger number of beam particles that do not interact with the target nuclei.

The discovery of the quadrupole and monopole vibrations in the 1970's was based on the development of two types of scientific instrument: machines capable of accelerating particles of various kinds to energies higher than 50 MeV and sensitive magnetic spectrometers (spectrographs) capable of separating particles of slightly different energies. Once the instruments were available nuclear vibrations began to be reported in rapid succession. The giant quadrupole vibration was first seen in electron-scattering measurements done at Darmstadt in West Germany by Rainer Pitthan and Th. Walcher and in proton-scattering measurements done at the Oak Ridge National Laboratory by F. E. Bertrand and M. B. Lewis. The giant quadrupole vibration has now been excited in practically all species of nuclei; it has a frequency that varies inversely with the radius of the nucleus. The equivalent energy is in the range from 10 to 20 MeV.

The first indications of the breathing mode were found with deuteron scattering by Nadine Marty and her collaborators at the Institute of Nuclear Physics of the University of Paris (Paris-Sud) at Orsay. Observation of the diffraction peak in the forward direction, providing definitive proof of the monopole mode, was first obtained in 1977 with alpha-particle scattering by a group at Texas A&M University led by Dave H. Youngblood. Like the quadrupole, the breathing mode has a frequency that decreases with increasing size of the nucleus. In energy units the range is from 15 to 25 MeV, slightly higher than the energy of the quadrupole. The giant octupole vibration was first detected in 1980 in scattering experiments done with 800-MeV protons at the Los Alamos Scientific Laboratory. The octupole mode has an energy ranging from 20 MeV in heavy nuclei to 30 MeV in lighter nuclei.

The history of theoretical speculation about nuclear vibrations and the interplay between theory and experiment



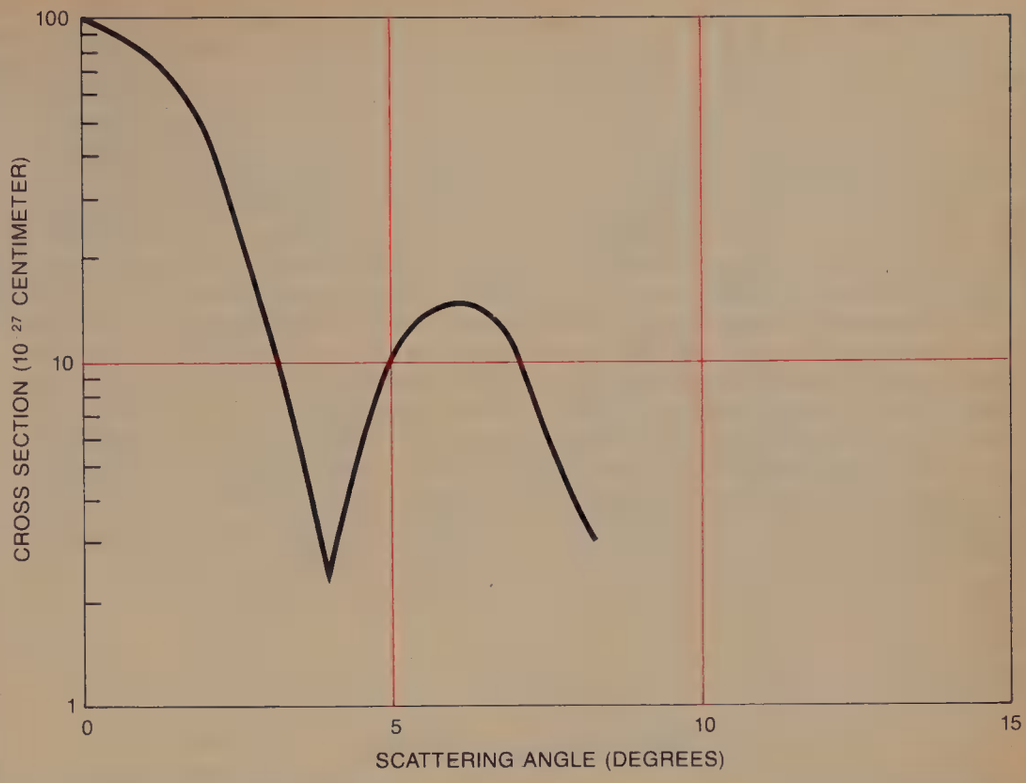
DIFFRACTION PATTERN associated with the monopole vibration is a point on the beam axis. The pattern is known as the Poisson spot; in the 19th century Siméon Denis Poisson proved mathematically that if light consists of waves, there should be a small bright spot in the shadow of a sphere. Poisson put forward his argument as proof that light cannot have a wave nature, since no such diffraction pattern had been seen at the time. The pattern is formed by a disk or a sphere whose surface absorbs light uniformly. Particles that interact uniformly with the surface of a nucleus, as do those that excite the monopole vibration, are diffracted with the same spot.

form an equally interesting thread in the story. Theoretical calculations predicted a nuclear phenomenon for which there was no experimental evidence, giving physicists some clues, albeit imperfect, about where to look and what it would look like if it were found. The experimental evidence in turn showed where the simplified theoretical models of vibrating nuclei were in error and helped to lay the groundwork for more powerful models based on more accurate assumptions.

The frequency of a vibrating body depends on two properties: the inertia of the constituent particles, which governs how quickly they respond to a force, and the restoring force, which opposes the displacement of the particles. The larger the inertia is, the slower the body vibrates, and the stronger the restoring force, the higher the vibrational frequency. More quantitatively, the square of the vibrational frequency is directly proportional to the strength of the restoring force and inversely proportional to the inertia.

In 1944 the Russian physicist Arkadii B. Migdal predicted the frequency of the dipole vibration by applying those simple mechanical laws to the nucleus. It is easy to construct a plausible model of the inertia of the nucleus vibrating in the dipole mode. Assuming that all the nucleons are moving with the same speed in the dipole vibration (but in different directions for protons and neutrons), the inertia is equal to the total mass of the nucleus. Modeling the restoring force is more difficult, and it is here the simplification in theoretical models of the nucleus generally comes in. In the case of the dipole mode the restoring force is mainly due to the attractive strong force between protons and neutrons. The strength of this force in a vibrating nucleus can be inferred from the binding energy of various nuclei. Among all nuclei those with nearly equal numbers of protons and neutrons have the greatest binding energy, apart from the effects of the protons' electric charge. A model of the proton-neutron interaction is constructed to fit the systematic variation in binding energy with nuclear composition; the same model is then used to calculate the restoring force when the nucleons are displaced in the dipole mode.

The measured frequency of the dipole vibration agrees remarkably well with the frequency Migdal predicted from his model. Nevertheless, we now know that more than the potential energy of separated protons and neutrons needs to be considered in modeling the restoring force of the dipole mode; furthermore, the inertial mass of the system is not simply that of free nucleons. The inertial mass is slightly smaller than Migdal assumed because in a quantum-mechanical description of the nucleus, protons and neutrons are not the only particles present. There are also pi mesons, or pions, the subatomic particles responsible for transmitting much of the strong interaction between nucleons. Pions are much lighter than nucleons and hence reduce the average inertial mass. For the dipole vibration, however, the discrepancies are minor and the predictions based on binding-energy calculations are essentially correct.



SCATTERING CROSS SECTION of nuclei vibrating in the monopole mode peaks on the beam axis (at zero degrees). The graph shows the angular distribution of alpha particles inelastically scattered from lead nuclei. The monopole vibration was identified seven years later than the quadrupole vibration, largely because of the difficulty of separating particles scattered along the beam axis from the particles passing through the target without being deflected.

Another model of the nucleus, the liquid-drop model, was proposed by Niels Bohr in 1936. Bohr took note of the fact that a nucleon in the interior of a nucleus is surrounded by other nucleons that pull it equally in all directions, so that the net force is zero. A nucleon at the surface, on the other hand, has other nucleons on only one side, and so it is pulled toward the center. The effect on the surface nucleons is analogous to the surface tension of a drop of water; in both cases the force tends to make the system take on a spherical shape.

The frequencies of liquid-drop vibrations were worked out by Lord Rayleigh at the end of the 19th century. In 1952 Niels Bohr's son, Aage Bohr, and Ben R. Mottelson suggested that the liquid-drop model might be applied to vibrations such as the quadrupole vibration, in which the nucleus oscillates between spherical and deformed shapes. In these vibrations the dominant restoring force should be the nuclear surface tension. Because the nuclear surface tension is relatively weak, Bohr and Mottelson predicted that the frequency of the quadrupole mode would be low. Quadrupole-shaped motions were found at low frequencies, but only a few of the nucleons, typically fewer than 10 percent, participate in the motions. The giant quadrupole vibration has a frequency much higher than the predicted one.

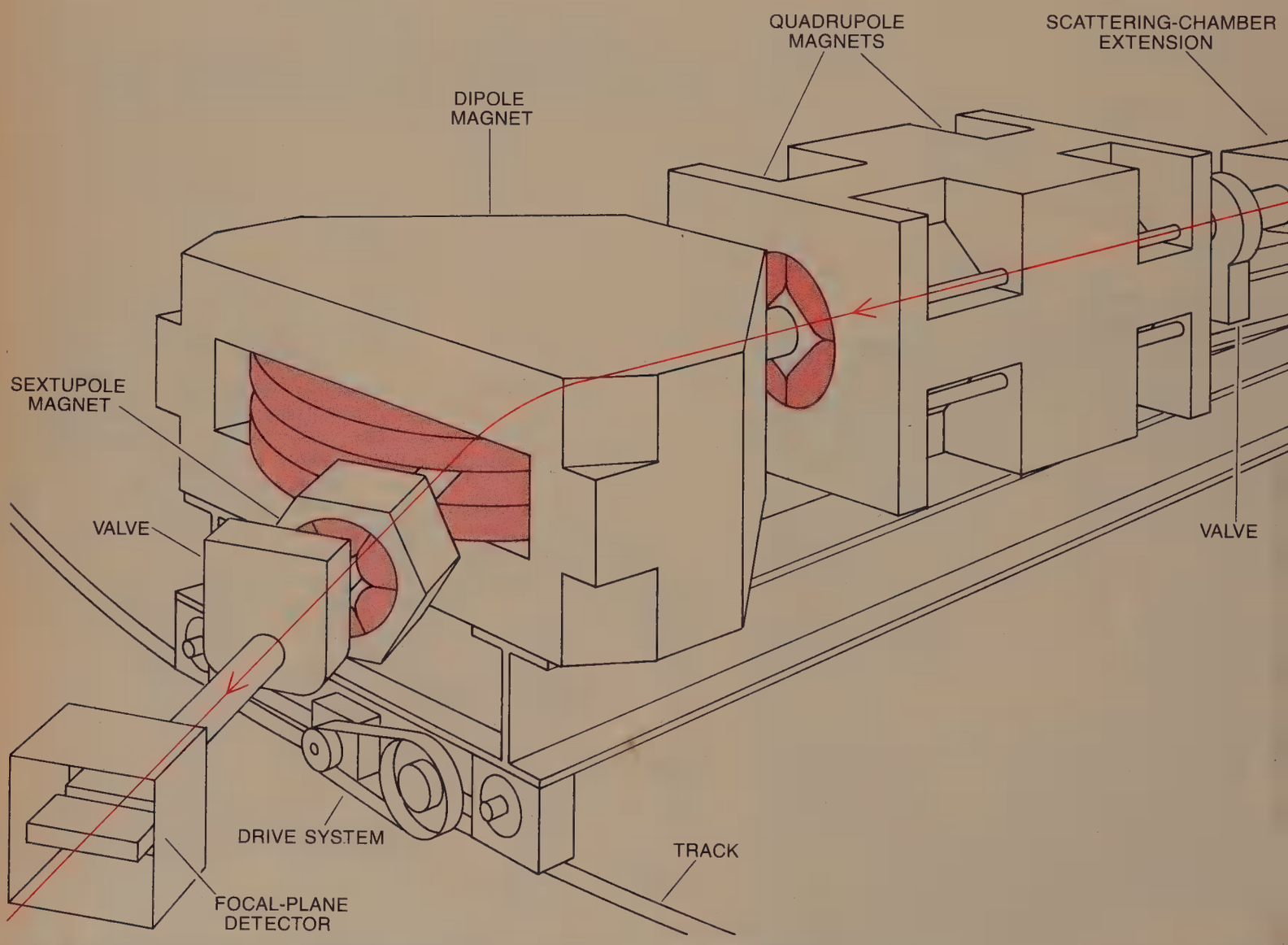
The problem with the liquid-drop model is that nucleons are not free to move in the nucleus in quite the same way as molecules are free to move in a drop of liquid. Nucleons must obey the Pauli exclusion principle, which states that no two identical nucleons can have exactly the same quantum-mechanical state of motion. As a consequence two protons or two neutrons having the same spin orientation must occupy different orbits in the nucleus. Their motion is thereby constrained to some extent; they must keep out of each other's way.

To include the quantum effects properly requires a much more elaborate theory, which I shall describe below. The theory gives a quite simple and unexpected result for the giant vibrations, namely that the nucleus has a rigidity making it respond more like a solid than a liquid. The restoring force for the quadrupole vibration is governed by the elastic constant of the nuclear medium. The value of the elastic constant can be estimated from the quantum theory; it is proportional to the kinetic energy of

the nucleons in their shell-model orbits. Similar physical principles determine the forces between ordinary atoms in a solid. If two atoms are pushed together, there is a repulsive force that can be traced to the electrons' increased kinetic energy, which in turn can be traced to the requirements of the exclusion principle. Nuclear matter is, however, much stiffer than ordinary matter because the nucleons in the nucleus have much higher kinetic energy than the electrons in the atom. Given the elastic constant of nuclear matter from quantum theory, the frequency of the quadrupole vibration can be calculated from a formula for the elastic vibrations of a sphere worked out more than a century ago by the British physicist A. E. H. Love.

The great stiffness of nuclear matter clearly suggests that nuclear quadrupole vibrations should be of high frequency. How, then, is one to understand the low-frequency quadrupolelike motions in which a few percent of the nucleons take part? If the disturbance that excites a vibration is slow, it is relatively easy to move a few of the nucleons into empty orbits of nearly the same energy. The exclusion principle does not forbid such a transition, and so the nucleus acts more like a liquid. When the nucleus is struck by a high-speed projectile, on the other hand, there is no time for internal rearrangement and the giant vibrations are more prominent. Overall the nuclear response to an external force resembles that of Silly Putty, which responds like an elastic solid to sudden forces and like a viscous liquid to slow ones.

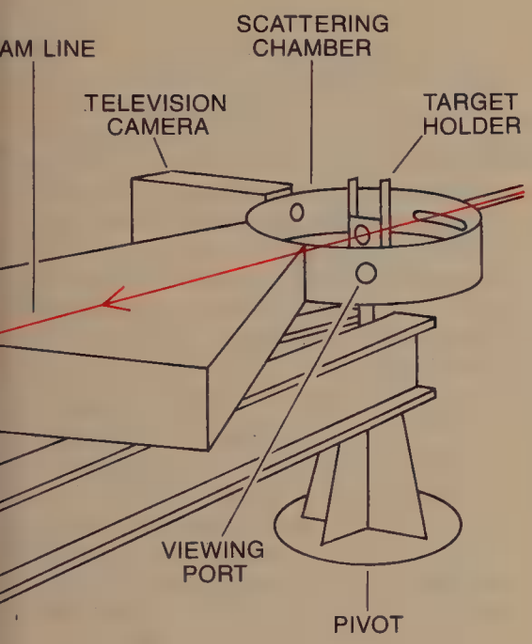
The frequency of the monopole vibration can also be worked out from a simple model of the nucleus in which only the inertia of the nucleons and the dominant restoring force of the vibration are considered. The restoring force in this case is the resistance of nuclear matter to compression. (In the "breathing" motion of a monopole vibration the nucleus is alternately compressed and rarefied.) The breathing model was suggested long before quantum calculations of the monopole vibration were available; later mean-field theory confirmed the



THE MAGNETIC SPECTROGRAPH, a sophisticated version of the mass spectrometer, separates particles that have lost a specific amount of energy in exciting a nuclear vibration from other particles in a beam. The beam of particles from an accelerator is directed onto a thin foil of target material in the scattering chamber of the spectrograph. After interacting with the target the projectiles are sorted according to energy by a dipole magnet that deflects them into a circular path. A particle with a higher energy follows a wider arc in the magnetic field than a particle with a lower energy. Only particles with the selected energy pass through the magnet to the detector. Quadrupole magnets focus the beam; a sextupole magnet corrects focusing aberrations. In the Michigan State University spectrograph shown here the detector is a multiwire gas ionization chamber. The high-energy particles ionize atoms of gas in the chamber. The electrons lib-

validity of a force based on the compressibility of nuclear matter. It has not been possible, however, to calculate the compressibility coefficient reliably or to measure it by other means, so that the observation of the monopole vibration provides the most direct information on the compressibility of nuclear matter. The coefficient derived from the breathing model together with the observed frequency of the vibration has nonetheless received spectacular confirmation from a very different source.

When a massive star comes to the end of its life, the inward-pushing gravitational forces are no longer balanced by the outward pressure of hot gases produced by nuclear reactions and the star begins to collapse. Gravitation compresses the core of the star to the density of nuclear matter, at which point the great resistance of nuclear matter to further compression begins to counteract the gravitational collapse. Depending on the compressibility of nuclear matter (among other things), the collapse may continue to the formation of a black hole or it may be stopped by an outward-moving shock wave that blows off the outer layers of the star in a supernova explosion. A dense neutron star is left at the center of the explosion; the maximum mass such a star can have depends directly on the compressibility coefficient. To date there is good agreement between the observed range of neutron-star masses and the compressibility coefficient deduced from the nuclear monopole vibration.



erated are attracted to the closest point on an array of parallel wires in the detector chamber. Currents induced in the wires are individually detected to determine the position of the particle. The spectrograph can be pivoted around the chamber so that the scattering probability can be measured as a function of angle.

An exact theory of nuclear structure (as opposed to a model) would specify in detail the forces exerted by each nucleon on every other nucleon. In a nucleus of moderate size, for example oxygen 16, there are 120 pairings of the nucleons to be considered. Furthermore, according to quantum mechanics, all the possible configurations of the nucleons have to be considered simultaneously, and each configuration must be assigned some amplitude in the quantum wave equation. The relative amplitudes would be independent of time in a description of the unexcited nucleus, but in the vibratory motion they would of course vary with time. In either case the task of describing the system in this way is mathematically intractable. The shell model offers a simplified description; it is approximate but nonetheless retains most of the quantum physics and is quite accurate in accounting for many of the properties of nuclei. The shell model does not attempt to calculate all the interactions of the individual nucleons. Instead a single potential, or mean field of force, is defined; it represents the collective effect of all the particles on any given particle. The quantum wave equation is solved for one particle at a time in this common potential.

The problem then becomes one of choosing the appropriate potential. The starting point is the distribution of particles in an unexcited nucleus; summing the fields associated with the individual nucleons yields an approximate collective potential. The next step is to go back and alter the wave functions of the individual particles in accordance with the estimated potential. By successive approximation one finds a potential and a set of particle wave functions that are mutually consistent. This method was introduced by William Hartree as a means of describing the electrons in an atom; the technique has been very useful in nuclear physics and is the basis of the nuclear shell model. The usual shell model is a static theory, but one can easily allow the potential field to depend on time in order to describe vibrations.

The time-dependent mean-field theory was first applied to nuclear vibrations in the 1960's. The change in the potential caused by an external force is calculated from the change in the distribution of the nucleons from one point in the vibrational cycle to the next. Again a repetitive procedure is employed to arrive at an accurate description of the field. The potential acting on the nucleons is calculated from their movement. The resulting approximation of the field in turn serves to refine the initial wave functions, which specify a new approximate potential. Ultimately a set of wave functions is found that fits both the distribution of particles and the potential field for the succeeding point in the vibrational cycle. The procedure can be simplified for small-amplitude vibrations.

The mean-field theory, rather than the simpler models of the nucleus, has provided the theoretical basis for vibrational studies for the past 10 years. In 1952 Mottelson predicted the frequency of the quadrupole vibration on the basis of the liquid-drop model; as I have stated, the predicted frequency turned out to be too low. In 1969 Mottelson predicted the frequency of the same vibration on the basis of the mean-field theory. When the vibration was finally detected in 1971, the observed frequency agreed with his second prediction.

The dipole, quadrupole and monopole vibrations are all geometric deformations with clear analogues in the vibrations of ordinary macroscopic bodies. The spin vibrations of the nucleus are quite different. Spin is an intrinsically quantum-mechanical property, and analogous motions are not to be found in macroscopic systems. The spin vibrations differ from the shape vibrations in that the spatial distribution of the nucleons may remain frozen with only the spin orientation varying with time. When there is no spatial motion, the Pauli principle is more restrictive and only a few of the nucleons can participate in the spin vibration. In the unexcited nucleus the spins are nearly all paired; the excitation process tips the spins of some of the nucleons and they precess. There are several ways the spin precession can be induced and detected. The nucleon spin has an associated magnetic moment and can interact through the electromagnetic field. Hence photon absorption and electron scattering are two techniques for studying spin properties of nuclei.

There are other fields as well that interact with nucleon spins. One of these is the field of the pi meson. When a pion

is absorbed by a nucleon, it changes the nucleon's spin orientation. An individual nucleon is surrounded by a pion field, so that even nucleon projectiles can induce spin vibrations. Another field that interacts with spin is the weak field, mediated by the recently observed W particle. The weak field is responsible for beta decay, one of the major processes in the formation of stable elements. In beta decay a proton is changed into a neutron or vice versa, usually with a reorientation of the spin of the affected particle.

One of the best experimental techniques for studying spin vibrations is the inelastic scattering of high-energy protons from a nuclear target. The pion field of the proton interacts with nucleons in the target, and much of the character of the interaction is due to the pion. If the exchanged pion is a neutral one, the electric charges of the projectile and the target remain the same and the collision is seen as an instance of ordinary inelastic scattering. It is also possible to exchange a charged pion, in which case the charges of the projectile and the target both change. The bombarding proton is turned into a neutron, and in the target one of the neutrons is changed into a proton. In spite of the charge exchange the same diffractive effects determine the angular distribution of the scattered beam. The simplest possible spin vibration, which is uniform over the entire surface of the nucleus, gives rise to a diffraction pattern with a peak at zero degrees, just as the giant monopole vibration does.

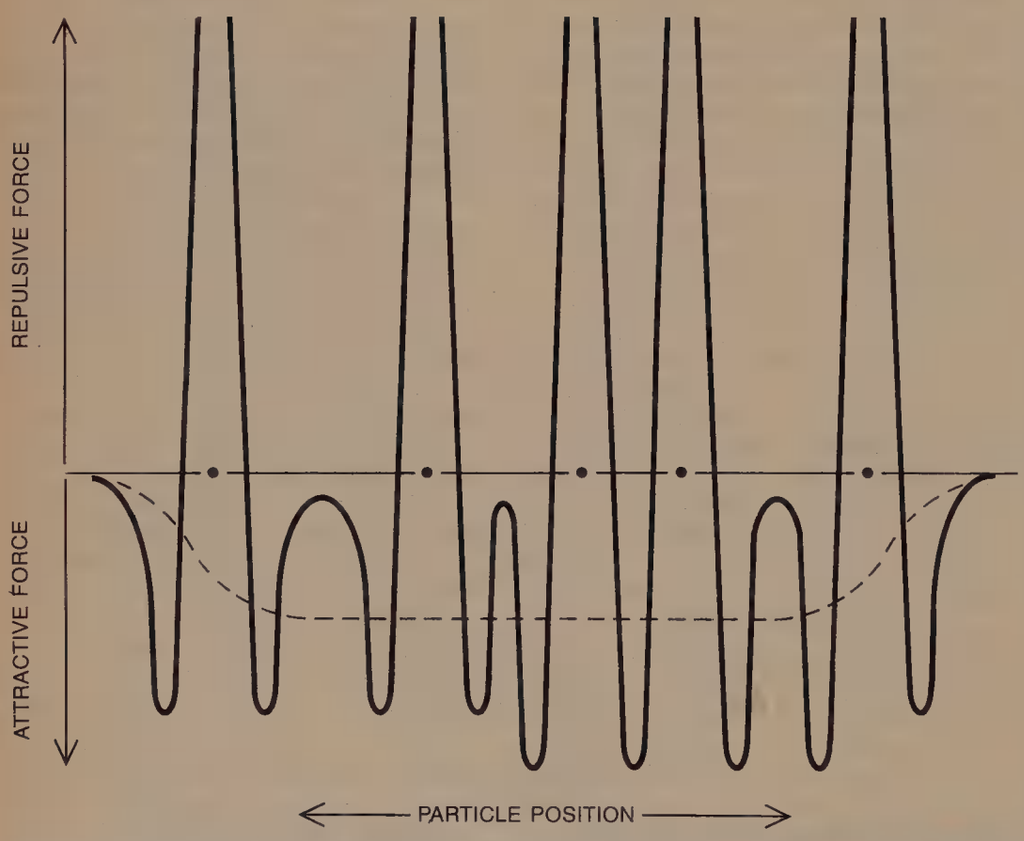
The charge-exchange spin vibration is called the giant Gamow-Teller resonance because of its relation to the spin-flip beta-decay process originally described by George Gamow and Edward Teller. The giant Gamow-Teller resonance was first observed in 1976 by Robert Doering, who was then a graduate student at Michigan State University. The energy of the proton beam was rather low and the forward diffraction peak was just barely distinguishable from the background of neutrons arising from more complex interactions with the target. More recent experiments done at the Indiana University Cyclotron Facility with higher-energy beams have given diffraction patterns in which the background is much reduced. The clearer patterns allow the properties of the vibration to be measured more accurately. The energy of the vibration lies in the range from 10 to 15 MeV.

The measured energies can be compared with theoretical predictions to check our understanding of the spin forces. There are no classical models for the spin vibrations, but the mean-field theory can be applied. The predictions agree well with the measurements, demonstrating that the theoretical description of the spin forces, based largely on the pion fields of the nucleons within the nucleus, is essentially correct. Even apart from theory, the existence of the giant Gamow-Teller resonance had long been suspected from indirect evidence. Beta-decay transition rates tend to be much lower than expected from the shell model, showing that the amount of spin precession at low frequency is small. From this finding one infers that the spin forces cause a higher-frequency precession—the giant vibration.

The last vibration I shall discuss is the magnetic-dipole resonance, which is much like the Gamow-Teller resonance. The main difference is that the number of protons and neutrons is not altered in the magnetic vibration. In the magnetic-dipole resonance the spin of the proton is tipped in the direction opposite to the spin of the neutron. Since the magnetic moments of protons and neutrons have opposite signs, the overall magnetic moment is maximized by this configuration. As the name of the vibration suggests, the magnetic field has a dipolar pattern.

It is relatively easy to study the magnetic-excitation properties of light nuclei because the structure is fairly simple when only a few nucleons are present. Depending on the nucleus and the degree of pairing, a prominent spin vibration may be present. Until recently, however, little was known about the magnetic excitations of heavy nuclei, in spite of a long search for the magnetic-dipole resonance. This situation has changed recently; in 1981 the magnetic-dipole resonance was found in the nucleus of zirconium 90 by using inelastic proton scattering to excite the spins of the target nucleons. The experiment was done at Orsay with 200-MeV protons and a spectrograph capable of identifying inelastic-scattering events at angles as small as two degrees. The excitation energy of the magnetic-dipole resonance ranges from 8 MeV in heavy nuclei to 15 MeV in light nuclei.

One interesting and somewhat puzzling aspect of both the Gamow-Teller and the magnetic-dipole resonances is that only a third of the expected number of nucleons seem to participate in the vibration in heavy nuclei. This is one of the reasons it was difficult to observe the magnetic-dipole resonance. The reduced strength has led nuclear physicists to look for a complete understand-



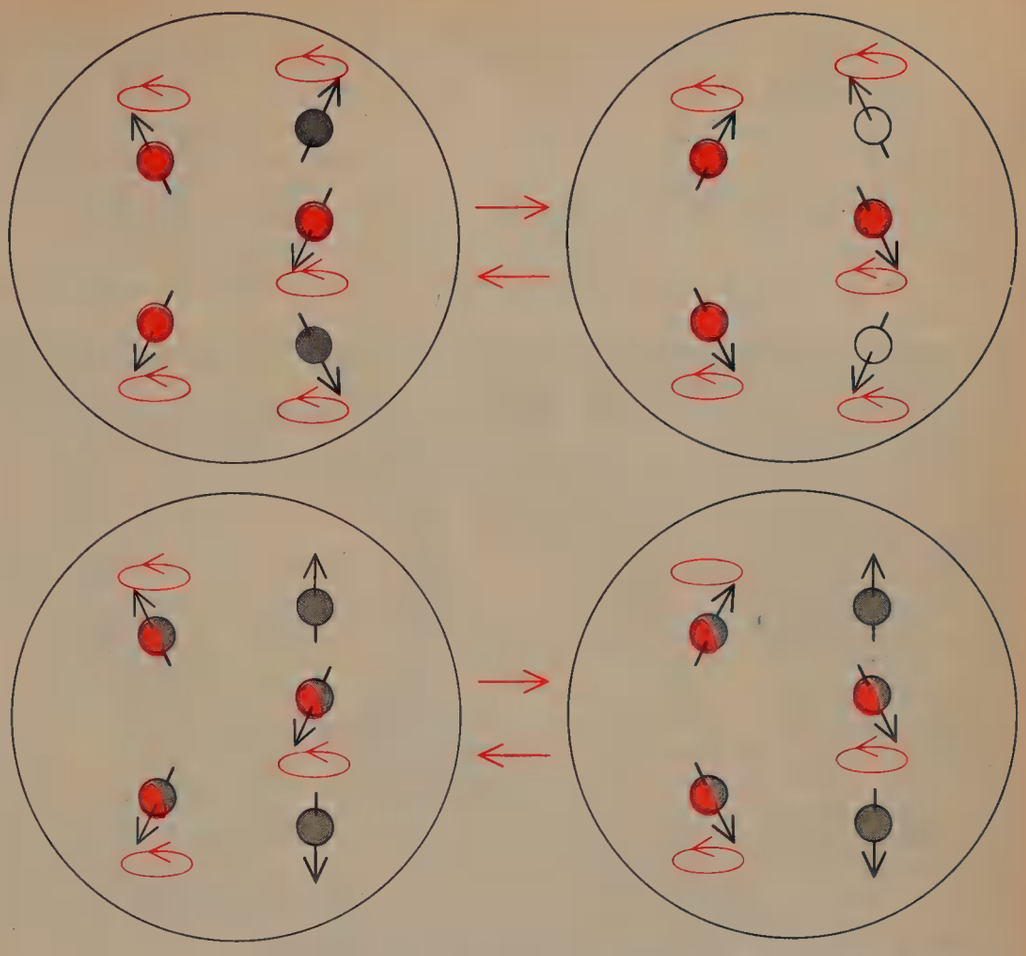
QUANTUM-MECHANICAL MODEL called the mean-field model provides a more satisfactory explanation of nuclear vibrations than models based on classical mechanics do. The force exerted on any one nucleon by the other nucleons in a nucleus fluctuates widely. Calculating the path of a nucleon in this field poses insurmountable difficulties. Not only is the field complex but also it changes rapidly because the nucleons creating the field are themselves in motion. The mean-field model simplifies the problem by postulating a smooth although varying field; it is the average of all the fluctuations of the actual field. An approximate mean field is calculated from the motion of the particles; the motion in turn is then calculated from the estimated field. Iterated calculations yield an approximation of the varying field that corresponds to the varying distribution of the nucleons during the vibration. The field shown here is for a one-dimensional array of nucleons; the actual distribution is a three-dimensional one.

ing of the vibrations on a deeper level. Protons and neutrons are no longer thought to be elementary particles; they are composite objects made up of the more fundamental particles called quarks. Any change in the nucleon's spin necessarily entails some change in the quarks' state of motion. The quark spins interact much more strongly than the nucleon spins and give the system a tendency to precess at a much higher frequency. The apparent number of spins vibrating at low frequency is thereby reduced.

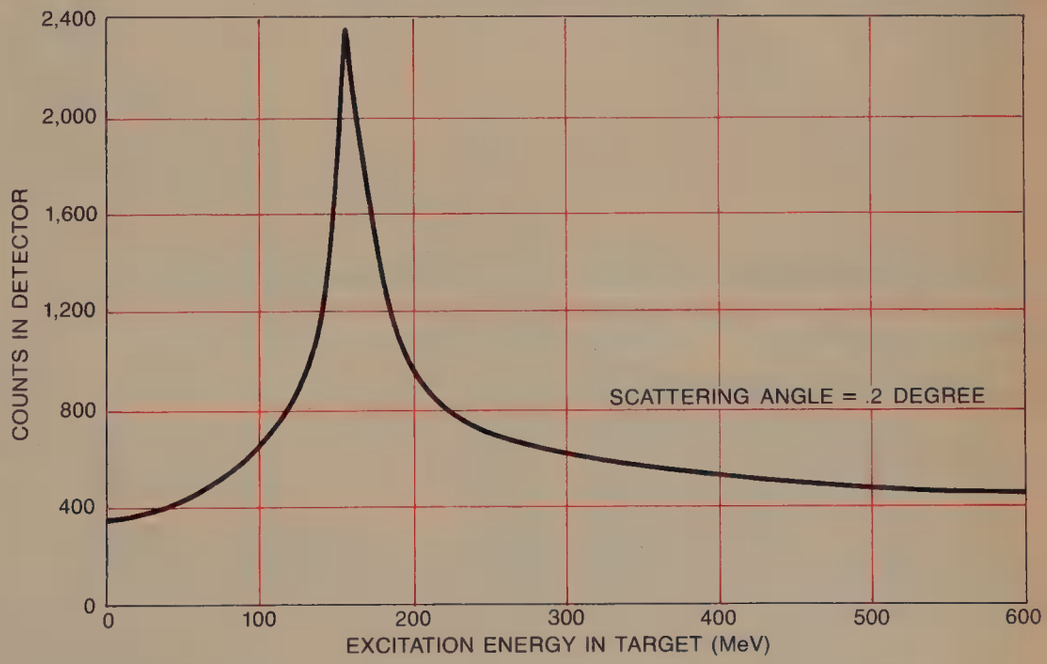
The reduced participation in the Gamow-Teller vibration has consequences for astrophysics, particularly for the theory of supernovas. During the collapse of a star the nuclei in the core undergo inverse beta decay in which a proton, an electron and an antineutrino combine to form a neutron. The amount of energy available to blow off the outer layers of the star turns out to be proportional to the fraction of the nuclei that take part in the process. Because the nuclear species involved do not exist in the laboratory, one must rely on the theory of the spin vibrations in seeking an understanding of how stars explode.

The vibrational models and the mean-field theory have successfully accounted for the frequencies of the vibrations, but they do not explain one important aspect of the motion: the damping of the vibrations, the decrease in the amplitude and the final extinction of the motion with time. Experimentally the damping shows up as a broadening of the resonance peak in measurements of the vibrational frequency. From the width of a resonance, say 4 MeV for a typical dipole or quadrupole vibration, it can be calculated that the nucleus oscillates through about three cycles before the motion is damped out.

The mechanisms that underlie damping are poorly understood. Collisions between nucleons would damp out the motion quickly, but the Pauli exclusion principle severely limits the probability of such collisions. Indeed, the mean-field theory would not work as well as it does if collisions were frequent. Another possibility is that the vibrational motion is coupled to more complex patterns of motion in the nucleus. The mean-field theory, however, is not capable of dealing with complex, uncoordinated motions of the nucleons. These motions are of course present and show up as the background cross sections in most measurements. The question of vibrational damping and the search for a theoretical model that can treat the more complex motions are now among the most active areas of research. No doubt there is much remaining to be learned about vibrations of nuclei, and indeed about the cooperative motion of quantum particles in general.



GIANT SPIN VIBRATIONS represent the coordinated precession of the spin axes of the nucleons. In the giant magnetic-dipole resonance (*top*) a proton and a neutron are tipped in opposite directions; their spin axes precess 180 degrees out of phase with each other. As a result of the coordinated precession of a fraction of the nucleons the nucleus as a whole acquires a net spin and a net magnetic moment. In the giant Gamow-Teller resonance (*bottom*) a proton is converted into a neutron and its spin is tipped from the original orientation.



SCATTERING CROSS SECTION of the giant spin vibration is much like that of the giant monopole vibration, peaking at an angle close to zero degrees from the beam axis. The difference is that the projectile, a proton, exchanges electric charge with a neutron in the target nucleus in the course of exciting the vibration, so that the scattered particles are neutrons rather than protons. In this case protons with an energy of 160 MeV excite the giant spin vibration in lead nuclei and yield a diffracted beam of neutrons with an energy of 146 MeV. Because the neutrons are uncharged, it is easy to separate them from the beam protons and to measure scattering in the direction of the beam. The neutrons are not deflected by a magnetic field, however, and so their energy cannot be measured in a magnetic spectrograph. Instead their velocity is measured by timing their flight to a detector placed 100 meters away from the target.

Microbodies in the Living Cell

A group of subcellular organelles linked by superficial structural similarities actually includes several distinct types, each having different sets of enzymes dedicated to different metabolic tasks

by Christian de Duve

In the early 1950's a Swedish graduate student, Johannes A. G. Rhodin, was exploring mouse kidney cells in the Karolinska Institute laboratory of Fritjof Sjöstrand, one of the pioneers of electron microscopy. He saw a small organelle that had not been described before. It was about half a micrometer in diameter (a micrometer is a thousandth of a millimeter), surrounded by a single membrane and filled with a fine granular matrix. The nondescript appearance of the unknown cell component particularly struck its young discoverer; for want of any characteristic feature to attach a name to, Rhodin (who is now at the University of South Florida College of Medicine) called it a microbody. Similar structures, many of them having a dense, semicrystalline core, were soon detected in liver cells. For several years there was no clue to the identity or function of the mysterious microbody.

Morphology reflects chemistry. By and large, all members of a given subcellular population (a distinctive organelle, for example) have the same enzymatic composition. By and large, each enzyme is found at a single site in the cell (in a particular organelle, for example). The same particle a cell anatomist sees in his micrographs may therefore reveal itself to the biochemist as a set of similar peaks in curves reflecting the distribution of particular enzymes in cell fractions separated by spinning in a high-speed centrifuge. It was therefore predictable that help in characterizing the microbody might come from biochemical studies. It did, but it came in the unpredictable, roundabout way that is common in basic research (to the confusion—would they only learn by it—of those who call for "goal-oriented" programming of scientific investigation).

In the mid-1950's my colleagues and I at the Catholic University of Louvain were in the midst of an exciting series of experiments (themselves prompted by a chance observation) that were just giving us a glimpse of a new intracellular organelle, the lysosome [see "The Lysosome," by Christian de Duve; SCIENTIFIC AMERICAN, May, 1963]. Our main tool was a modified cell-fractionation scheme. It had enabled us to demonstrate that in the liver of the rat the enzyme acid phosphatase is segregated within a special group of particles intermediate in size between mitochondria and "microsomes," the two main cytoplasmic entities recognized in those days. (Mitochondria are the power-generating organelles of the cell; microsomes, once thought to be organelles, are fragments of several internal cell-membrane systems.)

Applying our new fractionation scheme systematically to the enzymes other workers had found straddling the mitochondrial and microsomal fractions, we found that several acid hydrolases (enzymes that split a wide range of biological compounds in an acid medium) ended up along with acid phosphatase in our fractionations. It was the association of these enzymes that led us to the characterization of the lysosome: a saclike particle containing enzymes that digest nutrient molecules, foreign substances and, under certain circumstances, components of the cell itself.

A nondigestive, oxidizing enzyme showed the same distribution pattern, however. It was urate oxidase (or uricase, as it was then sometimes called): an enzyme that oxidizes urate (or uric acid) to allantoin and carbon dioxide. We had included it in our measurements because Alex B. Novikoff, who is now at the Albert Einstein College of Medicine, had reported that its centrifugal behavior was similar to that of acid phosphatase. We found the two enzymes did indeed move together when we fractionated liver tissue by differential centrifugation, which separates entities on the basis of the rate at which they sediment when they are spun in a centrifuge. This did not, however, necessarily mean that urate oxidase and acid phosphatase are associated physically within the same particle. It could instead reflect the cosedimentation of two distinct types of particles that happen to have the same sedimentation coefficient, which is a complex function of size, shape and density.

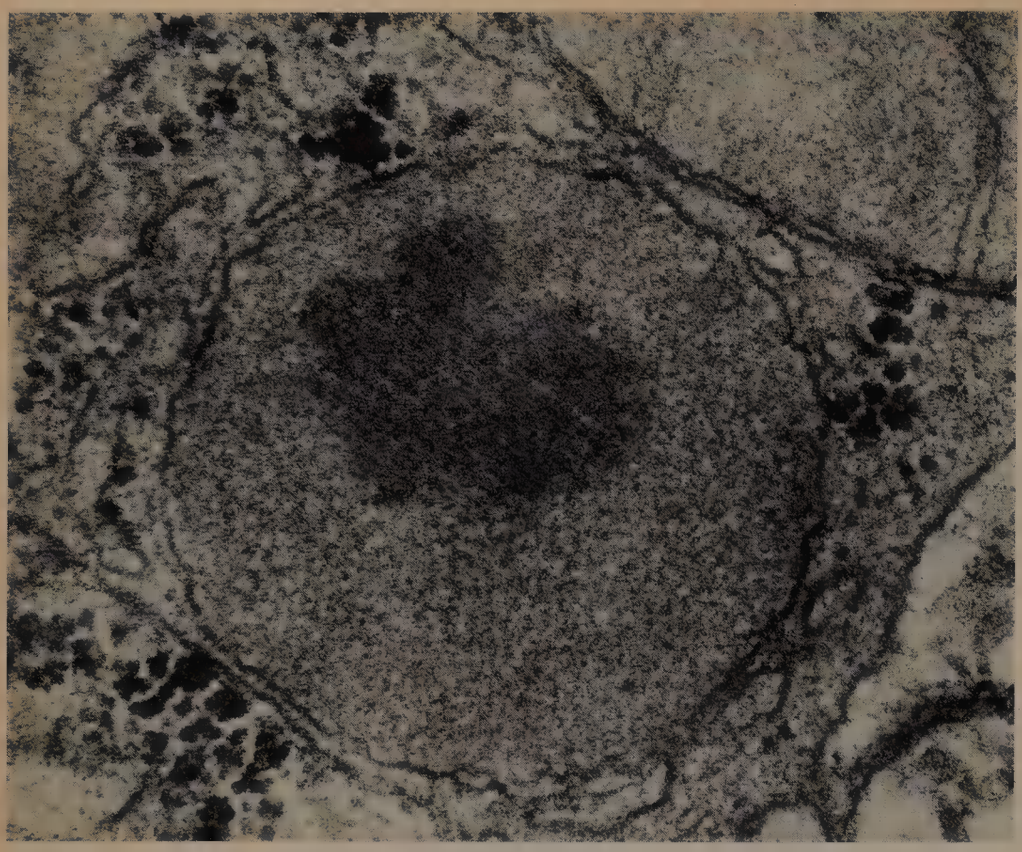
We suspected the second explanation might be true. One of our reasons was that urate oxidase did not (unlike acid phosphatase and the other acid hydrolases of lysosomes) display the property we called structure-linked latency: a seemingly inactive state of certain particle-linked enzymes that is observed when the particles are assayed for enzyme activity under conditions respecting their structural integrity. It was the chance discovery of this phenomenon (as exhibited by rat-liver acid phosphatase) that had attracted us to this field in the first place; we had been able to attribute it to the fact that the "inactive" enzyme is enclosed within a membranous envelope, inaccessible to the compound meant to serve as the enzyme's substrate in the assay. Our suspicion that urate oxidase is not in lysosomes was confirmed when we applied another fractionation procedure, density equilibration in a density gradient, in which each particle sediments only as far as its own density allows. In several systems of this kind acid phosphatase and urate oxidase showed clearly different distribution patterns.

The matter was clinched after my colleague Robert Wattiaux found that when rats are injected with the mild detergent Triton WR-1339, it accumulates in the lysosomes of liver cells and other cells and thereby causes a drastic and selective decrease in the density of those particles. The particles containing urate oxidase (and some other nondigestive enzymes that had been found to sediment with it) failed to show the change in density; they could therefore be separated cleanly from lysosomes, as well as from mitochondria, in a sucrose density gradient. This enabled my colleagues Pierre Baudhuin and Henri Beaufay to identify what we at first referred to as "uricosomes" (a term reflecting our ear-

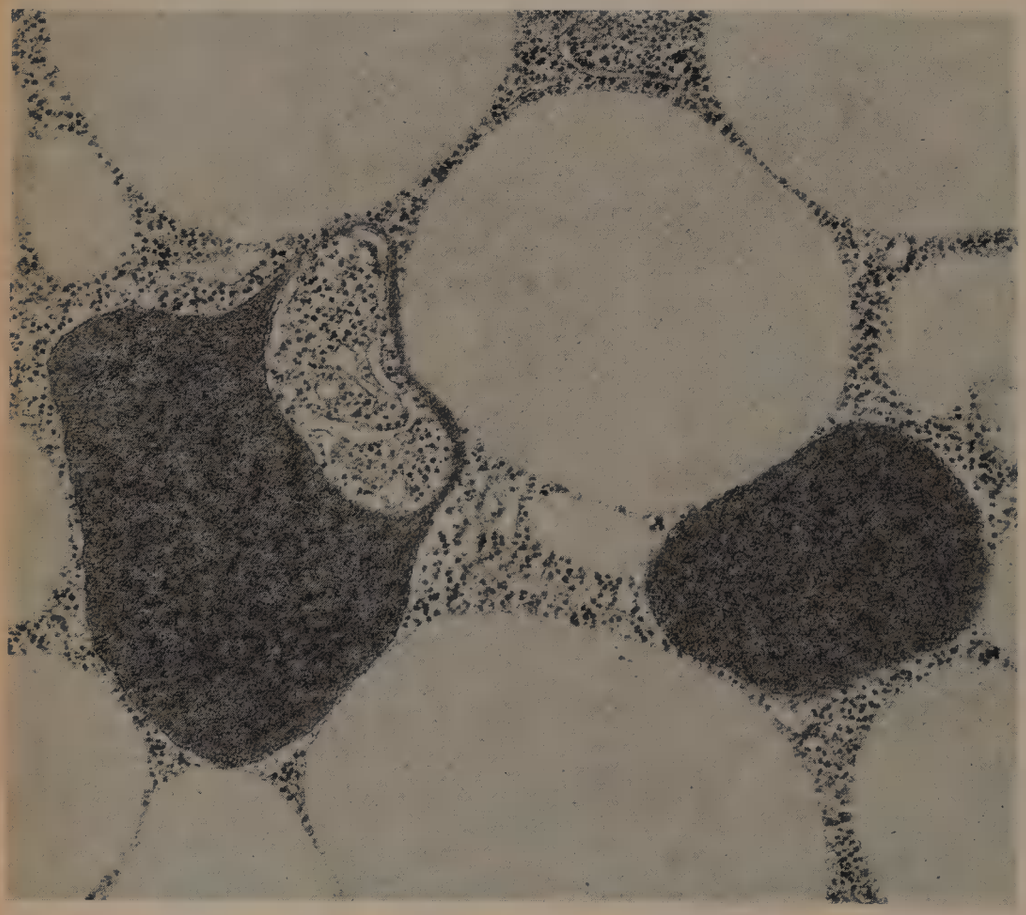
ly biochemical information) with the morphological entity Rhodin and others had described as microbodies.

The definition of the microbody kept changing as our search extended to enzymes other than urate oxidase and to biological materials other than rat liver. (The locus of our efforts also expanded. In 1962 I accepted a position at the Rockefeller Institute for Medical Research—now Rockefeller University—in New York. I did not sever my ties with Louvain, however, and have since divided my time between the two institutions. Something of a scientific airlift was soon established between the two laboratories, which continue to exchange investigators, findings and techniques and to collaborate on many projects.) First Baudhuin showed that catalase, an enzyme that breaks down hydrogen peroxide (H_2O_2), seemed to be associated with the rat-liver microbodies contain-

PEROXISOMES, the major type of microbody, are seen crammed into a yeast cell in this freeze-etch electron micrograph made by Marten Veenhuis of the State University at Groningen in the Netherlands. These are highly unusual peroxisomes whose proliferation has been induced by growing the yeast *Hansenula polymorpha* in a medium in which methanol is the source of carbon. The peroxisomes contain the enzymes alcohol oxidase and catalase, which respectively oxidize the methanol, generating hydrogen peroxide, and reduce the hydrogen peroxide to water. The yeast cells were frozen and then fractured. Ice on the fractured surface was sublimed, the surface was shadowed with platinum, and carbon was deposited to make a replica. Here the carbon replica of one cell is enlarged 36,000 diameters.



SINGLE RAT-LIVER PEROXISOME is enlarged 115,000 diameters in this electron micrograph made by Helen Shio in the author's laboratory at Rockefeller University. It is bounded by a single membrane and has a granular matrix with a denser, semicrystalline core. Part of a mitochondrion is visible (*top right*), as are sectioned channels of the endoplasmic reticulum.



TWO GLYOXYSOMES surrounded by lipid (fat) droplets in the cell of a germinating tomato seedling are enlarged 32,000 diameters in an electron micrograph made by Eldon H. Newcomb of the University of Wisconsin at Madison and Peter J. Gruber. One glyoxysome has an inclusion of cytoplasmic material. Glyoxysomes, another type of microbody, are in effect peroxisomes that contain some or all of the enzymes needed to catalyze the glyoxylate cycle, a variant of the Krebs cycle. Glyoxysomes of seedling cells also contain enzymes for what is called the beta oxidation of fatty acids; raw material for that oxidation is present in lipid droplets.

ing urate oxidase. Otto Z. Sellinger, who is now at the University of Michigan Mental Health Research Institute, did the same for D-amino acid oxidase, an enzyme that oxidizes the unusual D-amino acids found in certain bacterial constituents.

On a visit to New York, Baudhuin joined forces with Miklós Müller (who had come to us from Hungary with a major interest in the lysosomes of protozoa) and with a gifted graduate student, the late Brian H. Poole. The results of this collaboration were most instructive. Particles biochemically similar to the liver microbodies (except that they lack urate oxidase) were found in the mammalian kidney (where microbodies had first been detected morphologically) and also in the protozoan *Tetrahymena pyriformis,* a ciliate whose cytoplasm likewise contains microbodylike organelles. In addition to catalase and D-amino acid oxidase the particles were found to contain a hydroxy acid oxidase acting notably on glycolate and lactate and (in kidney but not in the other two materials) an L-amino acid oxidase acting on the usual amino acids that make up proteins.

These findings indicated we were dealing with a cellular organelle likely to be widely distributed in the animal kingdom. They also provided a possible clue to what the microbody does in life. We were particularly impressed with the fact that the various oxidases in the new particles, although they can oxidize a wide range of substrates, have one feature in common: in oxidizing these substrates they reduce oxygen to hydrogen peroxide. Catalase thereupon reduces the hydrogen peroxide to water. To reduce is to add electrons. For catalase the electron donors are such small molecules as methanol, ethanol, nitrite or formic acid or, in the absence of such a donor, hydrogen peroxide itself. Together the oxidases and catalase can therefore be pictured as forming a primitive respiratory chain in which electrons removed from various metabolites are eventually added to oxygen, forming water. Unlike the major respiratory chain of the mitochondria, this one is mediated by hydrogen peroxide and is not coupled to phosphorylations that yield adenosine triphosphate (ATP), the cell's major energy carrier.

On the basis of these considerations I proposed in 1965 that the new particles be named peroxisomes. The term has been widely accepted, although there is still something of a nomenclature problem. Be that as it may, the concept implied by the word peroxisome has provided a useful link connecting a variety of microbodylike particles that are found throughout the eukaryotic kingdoms.

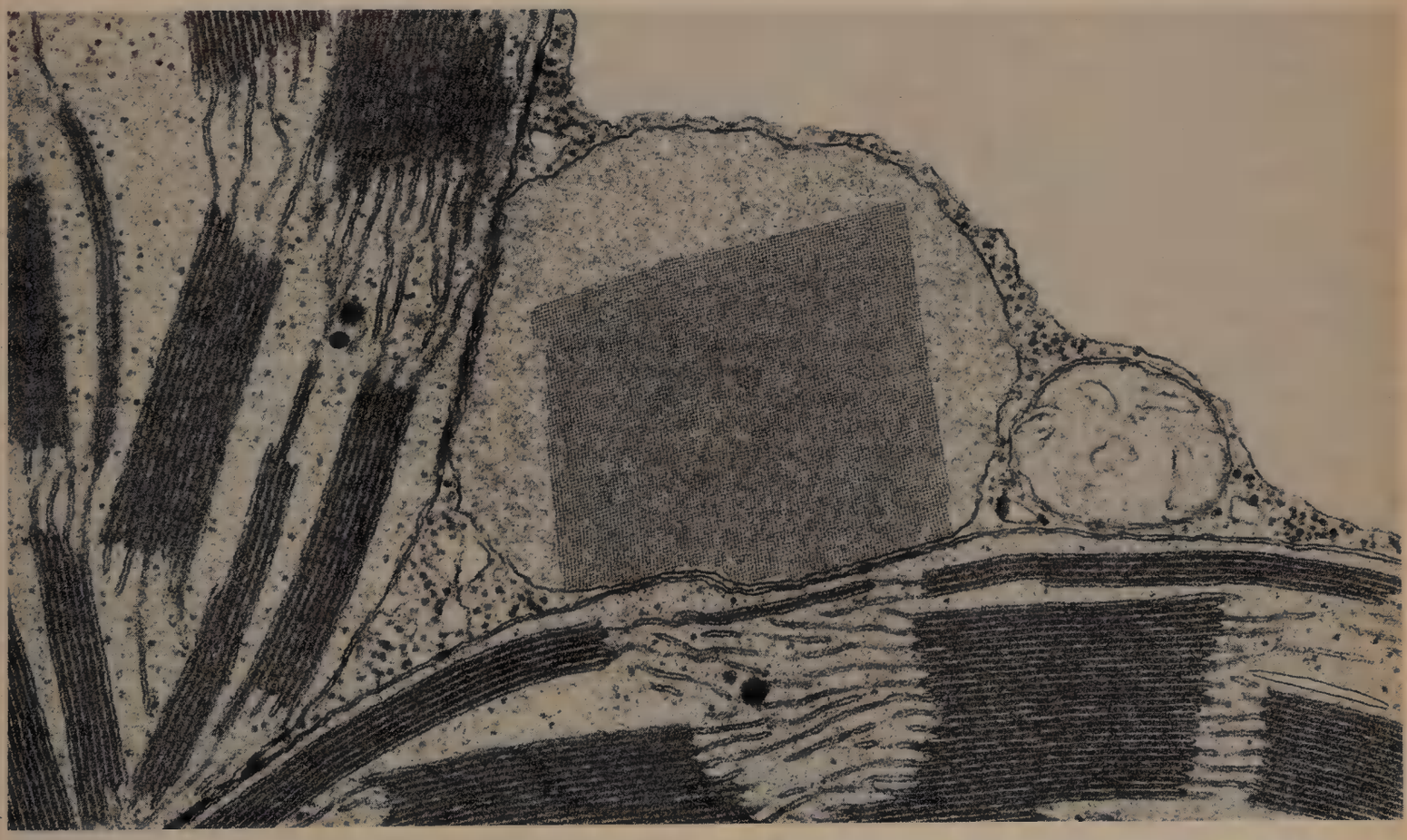
The concept is also physiologically significant, as has been shown most elegantly by Britton Chance and his coworkers at the Johnson Research Foundation in Philadelphia. By means of a sensitive spectrophotometric technique they measured the extent to which catalase is complexed with hydrogen peroxide in intact rat livers maintained in the laboratory. In agreement with the peroxisome concept they found that liver-cell catalase is normally complexed with hydrogen peroxide to be broken down; that the addition of known substrates for the oxidases increases the supply of hydrogen peroxide, and that the addition of substrates for the action of catalase accelerates the compensatory unloading of hydrogen peroxide from the enzyme-peroxide complex.

Implicit in the peroxisome concept is the central importance of catalase as the single agent of hydrogen peroxide disposal, whatever the nature of the oxidase substrates that have supported formation of the peroxide. Indeed, catalase seems to be present in all representatives of the peroxisome family, and often it is a major component of the particles; in rat liver, for instance, it accounts for as much as 15 percent of their total protein. A word of caution is required. Peroxisomes are certainly not the only site of hydrogen peroxide metabolism. There are oxidases that yield hydrogen peroxide in the cytosol (the fluid matrix of the cell), in the outer membrane of the mitochondrion and perhaps elsewhere in the cell. Most cells have at least one peroxidase in addition to catalase that is capable of breaking down hydrogen peroxide. There is also evidence that catalase itself is not always confined to the peroxisomes; there may be significant amounts in the cytosol.

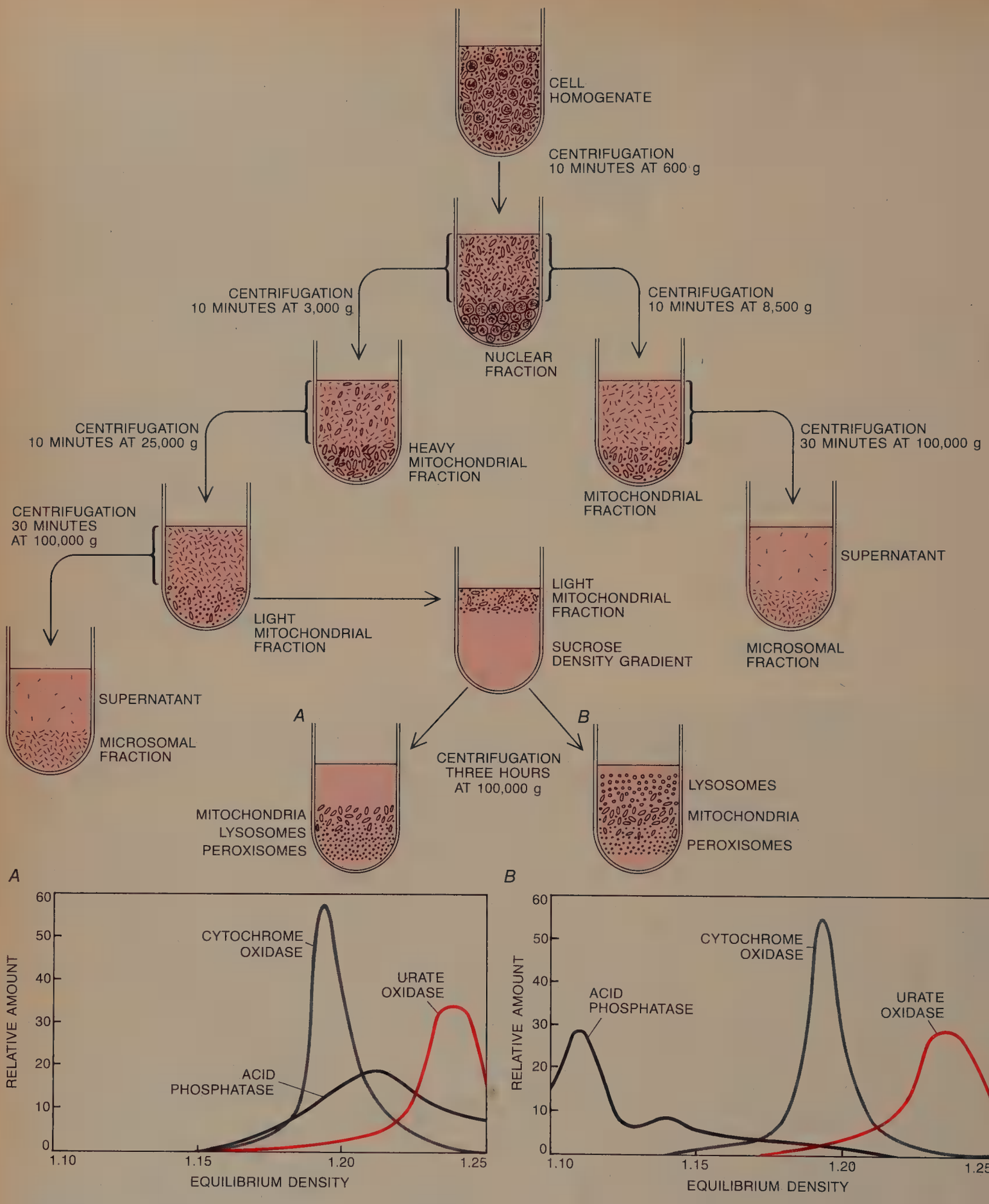
The next installment in the story begins with another transatlantic exchange. At the University of Oxford in 1957, in the laboratory of the late Sir Hans Krebs, his student Hans L. Kornberg discovered the glyoxylate cycle. In the 1930's Krebs had worked out the citric acid (or tricaboxylic acid, or Krebs) cycle, the final common metabolic pathway for the oxidation of fuel molecules. The glyoxylate cycle is a variant of the Krebs cycle. In it the sequence of four reactions whereby isocitrate is converted into malate in the Krebs cycle, with the release of two molecules of carbon dioxide, is bypassed; it is replaced by a two-reaction sequence that consumes an additional acetyl group furnished by acetyl coenzyme A (acetyl CoA) and leads to the formation of succinate as a by-product. Whereas one acetyl group is completely oxidized to carbon dioxide at every turn of the Krebs cycle, in the glyoxylate cycle (named for a key intermediate of the bypass) two acetyl groups are condensed to succinate at every turn.

Now, one of the major sources of acetyl CoA in living cells is what is called the beta oxidation of fatty acids. Moreover, the succinate generated from acetyl CoA by the glyoxylate cycle can be converted into carbohydrate by way of a segment of the Krebs cycle, phosphoenolpyruvate and reverse glycolysis. The glyoxylate cycle can therefore support gluconeogenesis, or the new formation of carbohydrate, from fat. It is the cycle's main biological function, and one of major importance.

Two workers attracted to Oxford by Kornberg's discovery subsequently took their interest to the U.S. One was Harry Beevers, who moved from Britain to Purdue University and set out to investigate the way in which the glyoxylate cycle is put in motion in germinating castor beans. This phenomenon, which is common to all fatty seedlings, serves to provide the young shoots with carbohydrate (at the expense of the seeds' oily stores) until the first green leaves appear and photosynthesis begins. Beevers and his co-worker Rowland W. Breidenbach located the two enzymes catalyzing the two glyoxylate-bypass steps in what ap-

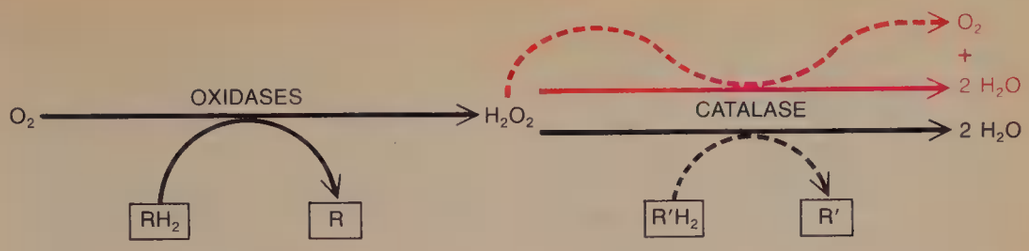


LEAF PEROXISOME, a large microbody housing a remarkably regular crystalline aggregate believed to consist of the enzyme catalase, is enlarged some 58,000 diameters in this micrograph made by Newcomb and Sue Ellen Frederick. The peroxisome is wedged between two chloroplasts and is adjacent to a mitochondrion in a tobacco-leaf cell; the large open area is the cell's central vacuole. The close association of the three kinds of organelle facilitates a circuitous metabolic pathway. Glycolate formed in chloroplasts is oxidized to glyoxylate in peroxisomes; subsequent steps in mitchondria and then back in peroxisomes lead to the formation of carbohydrate in chloroplasts.



CENTRIFUGAL FRACTIONATION separates cell components into fractions, which can be analyzed for their enzyme content to reveal the association of particular enzymes with particular cellular structures. A cell homogenate is spun in a centrifuge at progressively higher speeds. The classic fractionation scheme (*right*) yielded nuclear mitochondrial and microsomal fractions and a largely fluid supernatant. A modified procedure (*left*) developed in the author's laboratory some years ago subdivided the mitochondrial fraction. The resulting light mitochondrial fraction was layered onto a sucrose density gradient and subjected to density equilibration (*center*), separating its components into three layers on the basis of their density. The layers were defined by peaks in curves for three marker enzymes (*bottom*): cytochrome oxidase (a mitochondrial enzyme), acid phosphatase (known to be present in lysosomes) and urate oxidase, which appeared to be associated with a different particle. The peaks (and the fractions they define) were close together when normal rat cells were fractionated (*A*). When rats were injected with a detergent that accumulates in lysosomes, the density of the lysomes was selectively lowered (*B*), and they could be clearly separated from particles containing urate oxidase: the microbodies now known as peroxisomes.

peared to be a novel kind of cytoplasmic particle. The same particles contained the three other glyoxylate-cycle enzymes, which could also be found in the mitochondria as part of a regular Krebs cycle. The new particles could be purified by centrifugation in a sucrose gradient, where they equilibrated at a high density, as our liver peroxisomes had; they were found to have the general morphological characteristics of microbodies. In accordance with the new particles' biochemistry Beevers called them glyoxysomes.



PRIMITIVE RESPIRATORY CHAIN mediated by hydrogen peroxide (H_2O_2) is characteristic of peroxisomes and gives them their name. One of various oxidases in the particle oxidizes a substrate (RH_2), passing its electrons to oxygen to form hydrogen peroxide, which is reduced to water by catalase. Electrons for the reduction come either from one of various small molecules ($R'H_2$) or, if no other donor is available, from hydrogen peroxide itself (*color*).

Plant and animal biochemists are notoriously insulated from each other, and the connection between these findings and ours might long have gone unnoticed but for another former associate of Kornberg's. James F. Hogg of Queens College of the City University of New York had visited Oxford and there had begun investigating the glyoxylate cycle in *Tetrahymena pyriformis*. My colleague Müller, embarking on a search for lysosomes in the same organism, called on Hogg for technical advice and was easily convinced to cast his net wider: to look for glyoxylate-cycle and Krebs-cycle enzymes too. Their collaboration led to the discovery that the two glyoxylate-bypass enzymes are present in the peroxisomes of *T. pyriformis*. In the protozoan they are not, however, accompanied (as they are in castor-bean glyoxysomes) by the three enzymes the glyoxylate cycle has in common with the Krebs cycle; those enzymes are found only in the mitochondria with the other Krebs-cycle enzymes.

In the spring of 1967, at that great annual fair of American biological scientists known as the Federation meetings, Breidenbach announced the first findings on castor-bean glyoxysomes. Müller was there to report on the presence of glyoxylate-cycle enzymes in *Tetrahymena* peroxisomes. After the meeting it did not take the Purdue investigators long to detect in their castor-bean preparations both catalase and some oxidases (including urate oxidase) that yield hydrogen peroxide, thereby identifying glyoxysomes as peroxisomes. The name glyoxysome continues to take precedence over peroxisome whenever glyoxylate-cycle enzymes are recognized in the particle. This creates a certain amount of confusion because it is not exactly clear at what stage a peroxisome becomes a glyoxysome. Does it have to contain a complete glyoxylate cycle, as in fatty seedlings? Or are the two bypass enzymes enough, as in *T. pyriformis*? And what if only one of the two enzymes is found?

Apart from the nomenclature problem, the difference between the castor-bean glyoxysomes and the *Tetrahymena* peroxisomes is illuminating. In the seedlings the glyoxysomes contain a complete glyoxylate cycle and the mitochondria a complete Krebs cycle, so that there is duplication of the three enzymes common to both cycles. Gluconeogenesis and the oxidation are clearly separated, and each has its own supply of acetyl CoA. In *Tetrahymena* the mitochondria hold the whip hand by controlling the supply of isocitrate to the peroxisomes. Oxidation is clearly dominant, and only those isocitrate molecules that are not funneled into the Krebs cycle, and leak out of the mitochondria, are available to start down the gluconeogenesis pathway in the peroxisomes. (The mitochondria are needed for gluconeogenesis even in the fatty seedlings. Only there can succinate be converted into phosphoenolpyruvate to enter the reverse-glycolysis pathway.)

In higher animals the enzymes of the glyoxylate cycle appear to have been casualties of evolution. Mammals, including human beings, are unable to make carbohydrate from the common fatty acids. Recently, however, glyoxylate-cycle enzymes have been detected in cytoplasmic particles in epithelial cells of the bladder of the toad; they are said to be associated with catalase, as they are in the glyoxysomes (peroxisomes) of fatty seedlings and *T. pyriformis*. This finding supports the hypothesis that there is a phylogenetic relation between these particles in both the plant and the animal kingdoms.

In pursuing his glyoxysome studies Beevers, with T. G. Cooper, made the surprising discovery of another metabolic pathway: glyoxysomes in germinating castor-bean seedlings have all the enzymes needed for the beta oxidation of fatty acids to generate acetyl CoA, a capability thought to be limited to mitochondria. The topographical association of beta oxidation with the glyoxylate cycle makes for efficiency in the overall gluconeogenetic process. The glyoxysomal enzymes for beta oxidation catalyze the same reactions as their mitochondrial counterparts, with one interesting exception. The enzyme that initiates each cycle of oxidation, which in mitochondria is a dehydrogenase feeding electrons into the phosphorylating respiratory chain, is a simple hydrogen peroxide–producing oxidase in the glyoxysomes. Shades of peroxisomes!

It is astonishing that the wider significance of this observation failed to be recognized for almost 10 years (even by those who firmly believed plant glyoxysomes and animal peroxisomes are related phylogenetically) in spite of a highly suggestive finding made in the course of studies on clofibrate, a drug given to lower serum lipids in patients with a high blood-fat level. Investigators at the CIBA, Ltd., laboratory in Basel had found as early as 1965 that clofibrate induces a dramatic proliferation of peroxisomes in the rat liver.

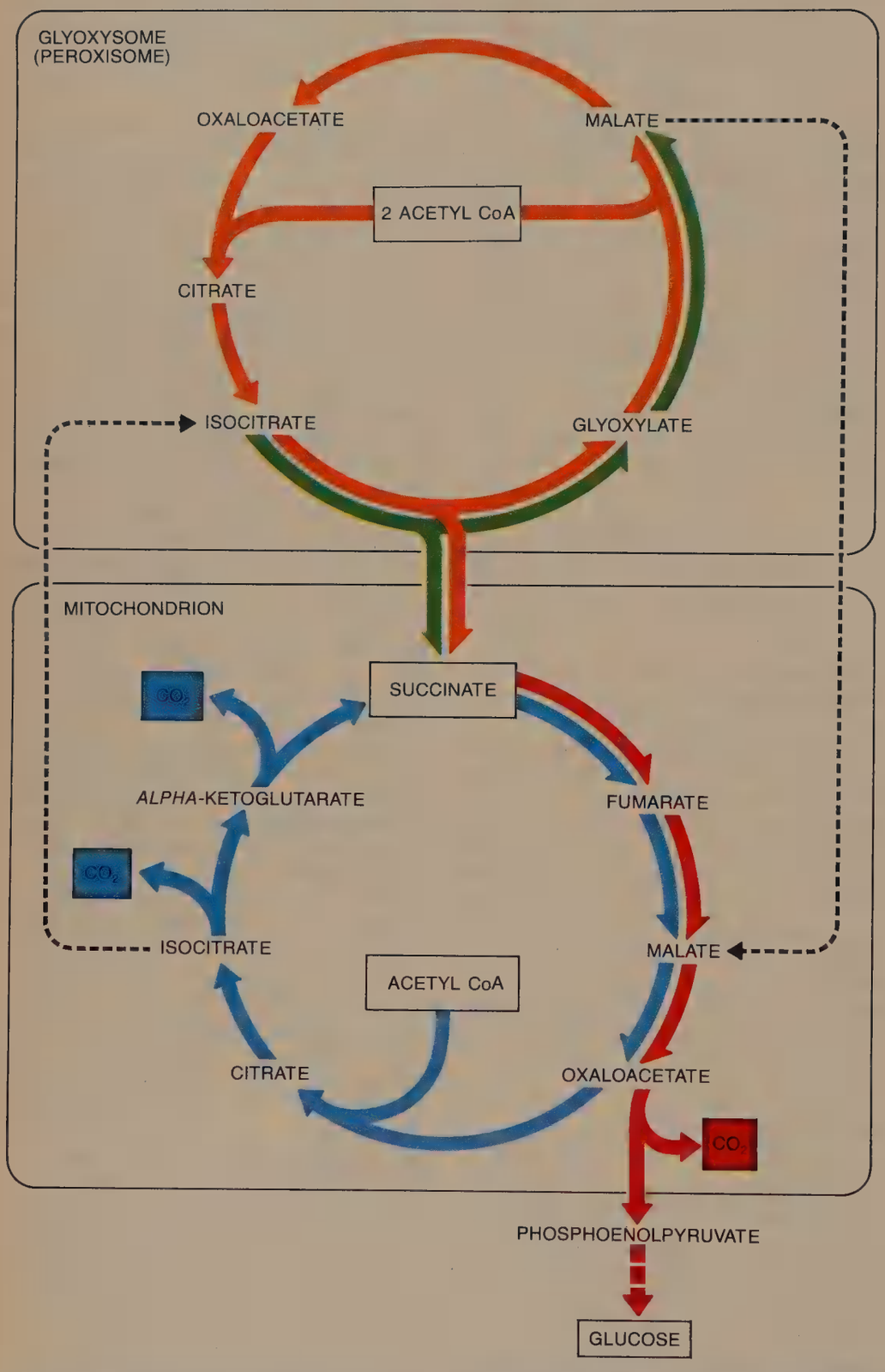
The observation attracted a great deal of interest and was confirmed elsewhere; other drugs that lower blood fats, even some unrelated structurally to clofibrate, were found to have the same effect, and the statement was repeatedly made that peroxisomes must play a role in lipid metabolism. Yet not until 1976 was this statement given experimental corroboration, when my associate at Rockefeller, Paul B. Lazarow, demonstrated that beta-oxidation enzymes are present in rat-liver peroxisomes and are enhanced by a factor of 10 or more after clofibrate treatment.

Lazarow's report elicited follow-up studies in a number of laboratories. The main peroxisomal beta-oxidation enzymes from rat liver were purified and characterized by Takashi Hashimoto of the Shinshu University Faculty of Medicine in Japan, who showed them to be entirely unrelated chemically to the corresponding mitochondrial enzymes. The peroxisomal beta-oxidation system has been induced experimentally with many different fat-lowering drugs; it can also be induced by high-fat diets or by exposure to cold, to both of which enhanced fatty-acid oxidation would seem to be an appropriate physiological response.

Imitating in reverse the role played by castor-bean glyoxysomes, liver peroxisomes mediated an important transdisciplinary fertilization process, this time with plant biochemistry as recipient and animal biochemistry as donor.

The result of the exchange was new information about photorespiration, a paradoxical phenomenon that has puzzled plant biochemists for decades.

As the term indicates, photorespiration is a process whereby light induces an increase in the amount of oxygen consumed by the tissues of green plants. It is associated with the production of glycolate by chloroplasts, the photosynthetic organelles. In the chloroplast ribulose-bisphosphate carboxydismutase, the key enzyme of carbon assimilation, has a dual function. In the absence of oxygen the enzyme acts exclusively to fix carbon dioxide, converting ribulose bisphosphate into two molecules of 3-phosphoglycerate. When oxygen is present, however, it competes with carbon dioxide; part of the ribulose bisphosphate is diverted into an alternative oxidative pathway that yields one molecule of phosphoglycolate and one of phosphoglycerate instead of two phosphoglycerates.



GLYOXYLATE CYCLE (*orange arrows*) **is a variant of the Krebs cycle** (*blue*), **the major metabolic pathway for oxidation of fuel molecules. In the glyoxylate cycle four Krebs-cycle steps are bypassed by a two-reaction sequence** (*green*), **which supplies succinate as a by-product. The succinate can be converted into glucose and other carbohydrates** (*red*). **Seedling-cell glyoxysomes have enzymes for a complete glyoxylate cycle; the Krebs-cycle enzymes are in the mitochondria, and so the enzymes catalyzing the three steps common to the two cycles are duplicated. In the protozoan** *Tetrahymena pyriformis* **the peroxisomes contain only the two bypass enzymes and lack the other three glyoxylate-cycle enzymes; the mitochondria supply isocitrate to the peroxisomes, which return malate to the mitochondria** (*broken black arrows*).

The subsequent breakdown of phosphoglycolate by a phosphatase yields glycolate, which is a major product of isolated chloroplasts and of certain unicellular algae when they are illuminated in the presence of oxygen. The larger the photorespiratory response, the smaller the net photosynthetic yield of a plant, and so the magnitude of photorespiration can have economic significance. The magnitude depends essentially on the ratio of oxygen to carbon dioxide, so that it varies inversely with a plant's ability to concentrate carbon dioxide. Tropical grasses such as corn, sugarcane and sorghum are particularly efficient in this respect.

The glycolate formed photooxidatively in chloroplasts was known to be further oxidized to glyoxylate, with the consumption of additional oxygen. Where and how this further oxidation takes place remained uncertain until a hydroxy acid oxidase acting on glycolate was found in association with catalase in the peroxisomes of rat liver. That finding stimulated N. Edward Tolbert of Michigan State University to overcome various difficulties besetting the homogenization and fractionation of green-leaf tissue. He thereupon discovered that peroxisomes are the site of glycolate oxidation. Like other representatives of this versatile family, leaf peroxisomes equilibrate at a high density in a sucrose gradient, and they display the morphological characteristics of microbodies.

These green-leaf microbodies, which have been studied particularly by Eldon H. Newcomb and his co-workers at the University of Wisconsin at Madison, are objects of great beauty. They may measure as much as 1.5 micrometers in diameter and often contain dense nucleoids, or crystalline inclusions, that are believed to consist of catalase. According to Tolbert, the glyoxylate made by oxidation of glycolate is converted into the amino acid glycine by an enzyme (a transaminase) in the peroxisomes. The scene then shifts to the mitochondria, where two glycine molecules are transformed oxidatively to the amino acid serine and carbon dioxide. The serine goes back to the peroxisomes, where it is converted into glyceric acid, which, after phosphorylation, can contribute to the formation of carbohydrate in the chloroplasts. This circuitous journey is reflected vividly in the morpho-

logical disposition of leaf peroxisomes: they are often seen wedged tightly between chloroplasts and mitochondria.

Microbodies containing peroxisomal and glyoxysomal marker enzymes were detected in the yeast *Saccharomyces cerevisiae* in 1968 by Charlotte J. Avers of Rutgers University. The study of yeast strains that can grow on unusual substrates suggests that their adaptability is due to certain remarkable inducible metabolic properties of their peroxisomes.

Saburo Fukui of Kyoto University and his colleagues first noted a marked proliferation of catalase-containing microbodies in strains of the yeast *Candida* that are grown on the long-chain hydrocarbons called alkanes. The particles contain catalase, several oxidases, a complete beta-oxidation system, the two glyoxylate-bypass enzymes (but not the three other enzymes of the glyoxylate cycle) and enzymes that remove hydrogen from long-chain alcohols and aldehydes. The mitochondria of the same organisms house the Krebs-cycle enzymes and alcohol dehydrogenases but no beta-oxidation system. These findings suggest that the peroxisomes (with the assistance of mitochondria to fill in the glyoxylate cycle) play a major role in making carbohydrates from long-chain alcohols that are derived from the alkanes.

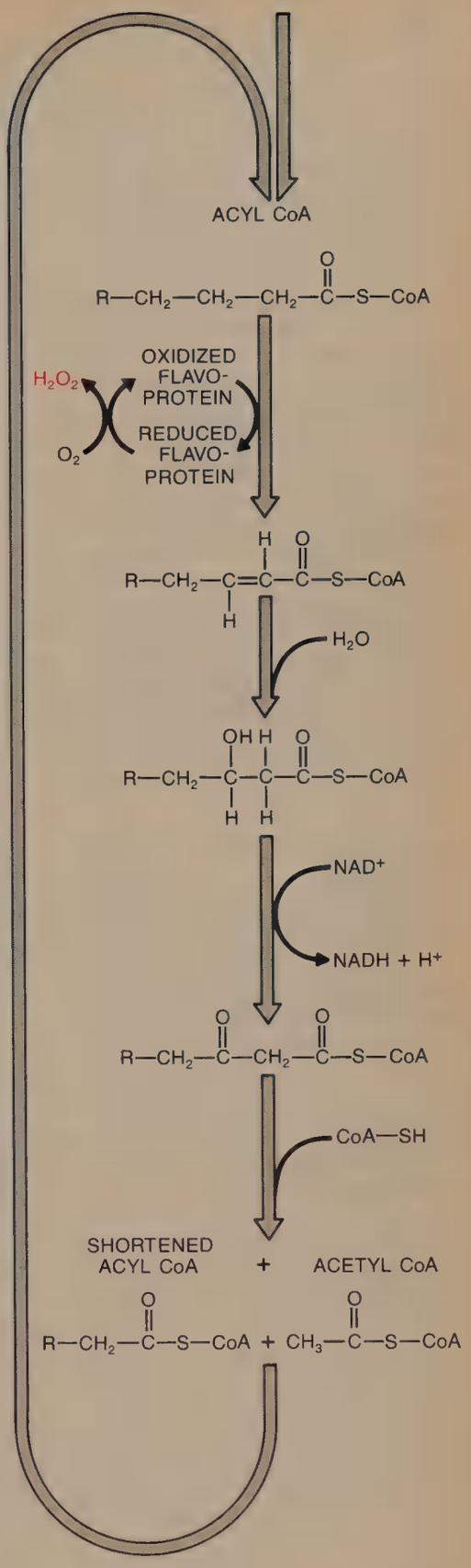
A most dramatic metabolic adaptation is seen when the yeast *Hansenula polymorpha* is grown on a medium in which the only source of carbon is methanol. Several groups of investigators in the Netherlands, Germany and Japan find such yeast cells are filled with large microbodies packed so tightly that they become almost cubic in shape. Much of the volume of these microbodies is occupied by a large crystalline structure composed of alcohol oxidase. The enzyme oxidizes the methanol to formaldehyde, generating hydrogen peroxide that is disposed of by catalase; the formaldehyde is subsequently converted into carbohydrate by a special pathway in the cytosol. There is evidence that yeasts may develop other peroxisome-mediated metabolic adaptations when they are grown on certain nitrogenous substrates such as D–amino acids, urate, choline or simple amines.

In mammals typical microbodies are seen only in a few organs, notably the liver and the kidney; these organs also have much higher catalase levels than other tissues. Small amounts of catalase are found, however, in every type of cell. When it is not in typical peroxisomes, the enzyme appears to be associated with smaller membrane-bounded particles, some .1 to .2 micrometer in diameter, which are often elongated and sometimes are even snakelike in form. These particles have been named microperoxisomes by Alex and Phyllis M. Novikoff and their collaborators. Increasing evidence for the presence of oxidases in these small structures suggests that they may indeed be related functionally to peroxisomes, even though they are different from microbodies morphologically.

How do peroxisomes (or glyoxysomes) originate? Do they resemble mitochondria, which are present in the germ cell and give rise to more mitochondria by growth and division? Or are they more like primary lysosomes, which arise from a pinching off of bits of the Golgi apparatus, which in turn is derived from the cell's major network of membranous channels, the endoplasmic reticulum? Many morphologists have reported seeing continuities between the membranes of plant or animal peroxisomes (or glyoxysomes) and the endoplasmic reticulum. These observations, together with some early reports that liver catalase is synthesized on ribosomes bound to endoplasmic reticulum (the "rough" endoplasmic reticulum), inspired the hypothesis that peroxisomes arise from the reticulum by budding. Peroxisomal proteins, it was proposed, are synthesized on membrane-bounded ribosomes, delivered into the lumen of the endoplasmic reticulum and thence transferred to peroxisomes by way of connecting channels.

In our Rockefeller laboratory first Poole and then Lazarow could find no evidence for this model. Lazarow and his collaborators find instead that proteins of the peroxisomal matrix are synthesized on free ribosomes in the cytosol. In experiments in which animals are fed amino acids labeled with a radioactive isotope, newly made (labeled) peroxisomal proteins are recovered entirely in the soluble fraction of homogenized liver cells. After a brief sojourn (an average of less than 20 minutes for catalase) in what one assumes (until there is proof to the contrary) to be the cytosol, the proteins are transferred to the peroxisomes by an unknown mechanism. Other investigators working on rat liver have reported similar results, and the main features of this scheme seem to apply also to the biogenesis of glyoxysomes in several different kinds of seeds.

The mechanism of translocation of the newly made proteins across the peroxisomal membrane does raise a puzzling problem. Proteins made by ribosomes on the rough endoplasmic reticulum are in effect injected into the membrane system by the ribosomes. In the case of proteins introduced from the cytosol into mitochondria, a "preprotein" chain is shortened as it passes through the mitochondrial membrane, perhaps providing energy for the transfer. In the absence of either active ri-



FATTY ACIDS are converted in seedling glyoxysomes to yield acetyl coenzyme A, the fuel for the glyoxylate cycle and hence for gluconeogenesis (the new production of glucose and other carbohydrates). The beta-oxidation pathway that accomplishes this conversion begins with an acyl coenzyme A, an activated fatty acid: a chain of CH_2 groups (some represented by *R*, for radical) linked by a CO group to a sulfur atom of coenzyme A. Each round of beta oxidation removes two CH_2 groups and yields a molecule of acetyl coenzyme A. In glyoxysomes (but not in mitochondria, the other site of beta oxidation) the first enzyme of cycle generates hydrogen peroxide, hallmark of peroxisomal metabolism.

bosomes to "push" large protein molecules such as enzymes into a membranous structure or proteolytic clipping to "pull" it in, where does the needed energy come from? One is tempted to think in terms of a self-assembly process within the peroxisome, with proteins that become attached to the particle membrane by a receptor-binding mechanism being somehow pulled through it and trapped inside the particle in multiprotein complexes. It is suggestive in this respect that dense aggregates, sometimes crystalline in nature, are so often observed inside peroxisomes.

Whatever the nature of the translocation mechanism, there is evidence that it can be reproduced in the test tube. When peroxisomal (or glyoxysomal) proteins, newly synthesized in the laboratory and labeled with a radioactive isotope, are mixed with peroxisomes (or glyoxysomes), some of the protein can be recovered from within the particles. This has been done with cucumber-seedling glyoxysomes by Wayne M. Becker of the University of Wisconsin and with rat-liver peroxisomes by Yukio Fujiki in Lazarow's laboratory.

The finding that proteins in the peroxisome's matrix are translocated from the cytosol, does not in itself rule out the possibility that the particle's membrane may originate from the endoplasmic reticulum and may even remain attached to it. It is equally possible, as Lazarow has pointed out, that peroxisomal-membrane proteins too are synthesized on free ribosomes, and that they are then inserted into the membrane of preexisting peroxisomes. Mitochondrial-membrane proteins and even some proteins of the endoplasmic reticulum originate in this way.

The fact that the Novikoffs and others have described direct continuities between peroxisomes and the endoplasmic reticulum would seem to argue in favor of the first possibility, were it not for the difficulty of identifying endoplasmic reticulum unequivocally on purely morphological grounds. Lazarow and Helen Shio have looked for continuities by staining for specific products of peroxisomes and of the endoplasmic reticulum. In no case were such products seen en route from one structure to the other; there is no luminal communication between the two. Lazarow has also shown, with Fujiki and others, that the membranes of the two structures have a very different chemical composition.

In view of these facts one should consider the possibility that what looks like a system of connections to the endoplasmic reticulum may conceivably be a system linking peroxisomes into some kind of special "peroxisomal reticulum" (Lazarow's phrase) or "microbodial compartment" (a term used by K. Gorgas of the University of Heidelberg). Evidence for such a compartment in at least certain cell types has come from workers in Germany and Portugal who have combined staining with serial sectioning. Such reconstruction has not yet been done for liver cells, but the characteristic clustering of liver peroxisomes and their striking biochemical homogeneity have long suggested to us that there too microbodies may be interconnected. It remains for future work to clarify this important issue.

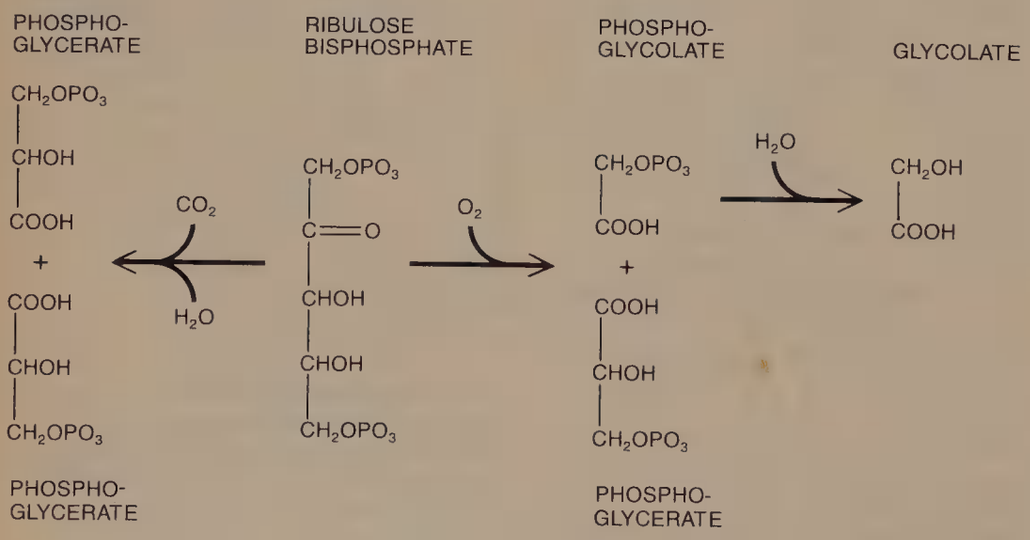
Particles identified biochemically as peroxisomes or glyoxysomes have been found in a wide variety of cells in all kinds of organisms above the bacteria. Whether these particles belong to plants, fungi, protozoa, invertebrates or higher animals, they show so many similarities and overlaps of all kinds that one can consider them members of a single family. For the sake of simplicity I shall refer to them henceforth as peroxisomes. My reason is that glyoxysomes, whether complete or incomplete, share the important properties of peroxisomes, whereas the reverse is not always true. To call them microbodies, on the other hand, could be misleading. As will become clear, not all microbodies are peroxisomes.

Peroxisomes in different organisms and tissues harbor various overlapping sets of enzymes from a growing list that already includes more than 40 distinct enzymes. So far no peroxisome has been found that has all the enzymes on the list; whether this is owing to the absence of certain genes or merely to their repression in a particular species is a question worth raising, given the striking degree of inducibility manifested by many peroxisomal enzymes.

For all of the functional diversity of peroxisomes, two metabolic features emerge as constants. One is the association of hydrogen peroxide–producing oxidases with catalase, the characteristic to which peroxisomes owe their name. The substrates of this oxidative activity are numerous and varied. They include amino acids (both D and L), hydroxy acids, fatty acids, alcohols, amines and a purine, but not the major electron carrier reduced nicotinamide adenine dinucleotide (NADH). Apparently any electrons delivered to NAD^+ (the oxidized form of the carrier) inside the peroxisomes have to be shuttled to the mitochondria for oxidative disposal. Alternatively, they could be handed over in the cytosol to a potential peroxisomal substrate such as lactate or ethanol and so be returned to the peroxisomes for final disposal. No energy-retrieval mechanism has yet been found to be coupled to peroxisomal oxidations.

The second function that comes through as some kind of metabolic leitmotif of peroxisomes is gluconeogenesis. I have mentioned its occurrence in germinating fatty seedlings, in *Tetrahymena,* in the leaves of photorespiring plants and in yeasts growing on alkanes or methanol. Even in mammals one is struck, as we were many years ago, by the fact that peroxisomes are seen mainly in the two major sites of gluconeogenesis, the liver and the kidney, but so far there is no evidence that they participate in the new formation of carbohydrate in these tissues.

The wide distribution of peroxisomes in all major groups of eukaryotes (or-



PHOTORESPIRATION competes with photosynthesis in the chloroplast. In photosynthesis carbon dioxide is fixed by an enzyme that converts ribulose bisphosphate into two molecules of phosphoglycerate (*arrow to left*). In the presence of oxygen the same enzyme acts as an oxygenase, diverting some of the ribulose bisphosphate into a pathway yielding only one molecule of phosphoglycerate along with a molecule of phosphoglycolate (*arrow to right*). The hydrolysis of phosphoglycolate produces glycolate. The glycolate is exported to peroxisomes and oxidized to glyoxylate, which is converted into glycine. Two glycine molecules are transformed in the mitochondria to make serine and carbon dioxide. Back again in peroxisomes, the serine is converted into glyceric acid, which contributes to carbohydrate formation in chloroplasts.

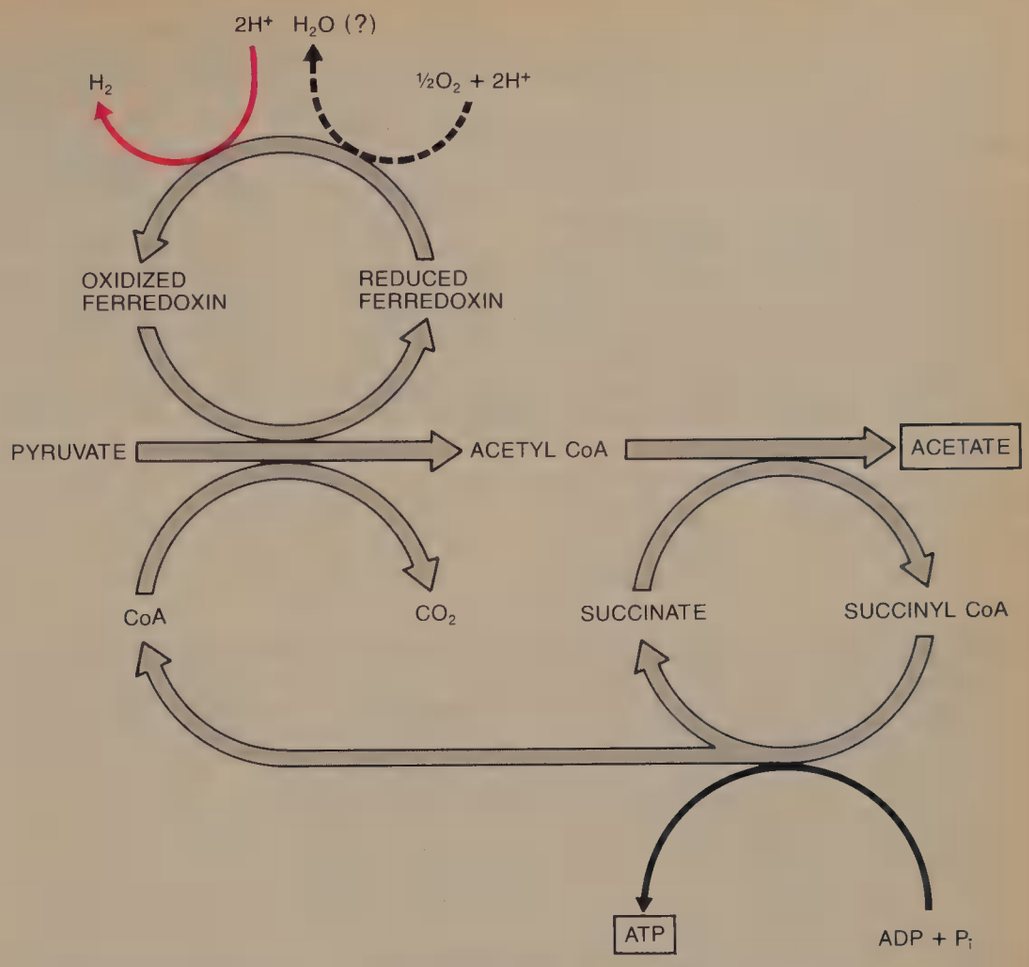
ganisms with nucleated cells) indicates that the organelles existed in the primitive unicellular organism thought to be the common ancestor of all eukaryotes. The rudimentary character and energetic wastefulness of the peroxisome's respiratory apparatus suggest that peroxisomes preceded mitochondria in the ancestral organism.

If peroxisomes supported oxidative metabolism in some primitive cell devoid of mitochondria, the possibility exists that the original cell line may have survived to this day. This idea prompted Müller to look for protozoa that have no mitochondria but do have microbodies. He chose the trichomonatids, a group of flagellates that, in addition to being endowed with beautiful microbodies and no mitochondria, have great economic and medical importance as parasites of the genital tract in animals and human beings and as agents of sexually transmitted diseases.

With Donald G. Lindmark, who is now at Cornell University, Müller succeeded in isolating and characterizing the trichomonad microbodies. He did not find what he was looking for, however. Instead of descendants of an ancestral peroxisome he found hydrogenosomes: organelles capable of oxidizing pyruvate to acetate and carbon dioxide. The oxidation is coupled to the synthesis of ATP. The electrons released by the oxidation are transferred to oxygen if it is available; under anaerobic conditions the electrons are transferred instead to protons (H^+) to form hydrogen (H_2).

Hydrogenosomes owe their ability to produce hydrogen to their having both the appropriate hydrogenase and a special electron-transport protein (a ferredoxin with a low oxidation-reduction potential) that serves as the electron acceptor for the enzyme pyruvate dehydrogenase. The ferredoxin also enables hydrogenosomes to reduce certain low-potential nitroimidazole derivatives and thereby convert them into compounds that are highly toxic to the cell. Such a property can be turned against its owner, and nitroimidazole drugs have become the standard treatment for trichomonad infections. Natural selection has not failed to wage its counteroffensive, and Müller is now tracking down pathogenic strains that have become resistant to the available drugs.

Quite aside from its considerable biological interest, the discovery of hydrogenosomes thus contributes to the war against what has become the most widespread venereal disease in the U.S. And once again I must point out to the funders and planners of scientific research that Müller was not proposing to enroll in this public-health cause—or even trying to understand hydrogen production in trichomonads. He was angling for an ancient peroxisome and came up instead with a hydrogenosome at the end of his line. So far no biochemical kinship has been found between the two types of microbodylike particles. Presumably they are not related.



UNUSUAL MICROBODY, the hydrogenosome, was discovered in *Tritrichomonas foetus*, a protozoan parasite that lacks mitochondria. Hydrogenosomes house enzymes that oxidize pyruvate, in the process generating adenosine triphosphate (ATP), the cellular energy carrier. Electrons removed from pyruvate are transferred to the electron-transport protein ferredoxin and finally to oxygen if it is available (*broken black arrow*), possibly forming water. In absence of oxygen electrons are transferred to protons (H^+), generating molecular hydrogen (*colored arrow*).

The latest (but perhaps not the last) of the microbodies was discovered by Fred R. Opperdoes in the laboratory of Piet Borst at the University of Amsterdam. Behind its perfectly legitimate morphological credentials this microbody turned out to be hiding a large segment of the glycolytic chain, the major pathway for the breakdown of glucose into smaller carbon compounds. A number of protozoan blood parasites (including the agents of African sleeping sickness, of the dreaded Chagas' disease of Latin America and of leishmaniasis in many parts of the world) have been found to harbor this unusual organelle. It has been named the glycosome.

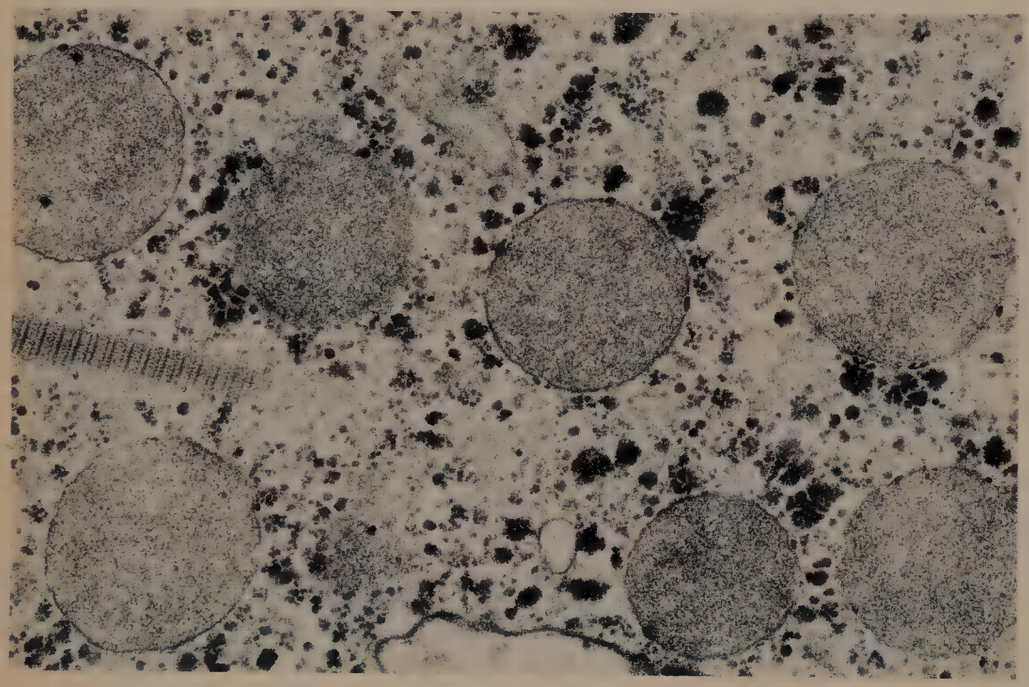
In all other animal cells so far investigated in this respect, the enzymes of the glycolytic chain are free in the cytosol; to find much of this system sequestered within a membrane-limited particle is very peculiar indeed. Aerobically glycosomes collaborate with mitochondria in supporting the energy needs of the cells in which they are found. They can also provide energy on their own in the absence of oxygen or if the mitochondrial oxidative pathway is blocked by the drug salicylhydroxamic acid. Their enzymes have physicochemical properties quite unlike those of the usual glycolytic enzymes, even though they catalyze most of the same reactions. Glycosomes may therefore be a vulnerable target for chemicals that interfere selectively with their special enzymes; such chemicals could serve as drugs to combat the deadly diseases caused by glycosome-carrying parasites. (Again basic research has practical application. Opperdoes now heads the tropical-disease unit at the International Institute of Cellular and Molecular Pathology, which we have established in Brussels specifically to facilitate this kind of fertilization.) Glycosomes, for all of their morphological similarity to the microbodies of animal and plant cells, are very different biochemically from peroxisomes and are probably not related to them.

A seemingly simple and still widely accepted explanation for the evolutionary origin of microbodies is to view them as offshoots of the main cytomem-

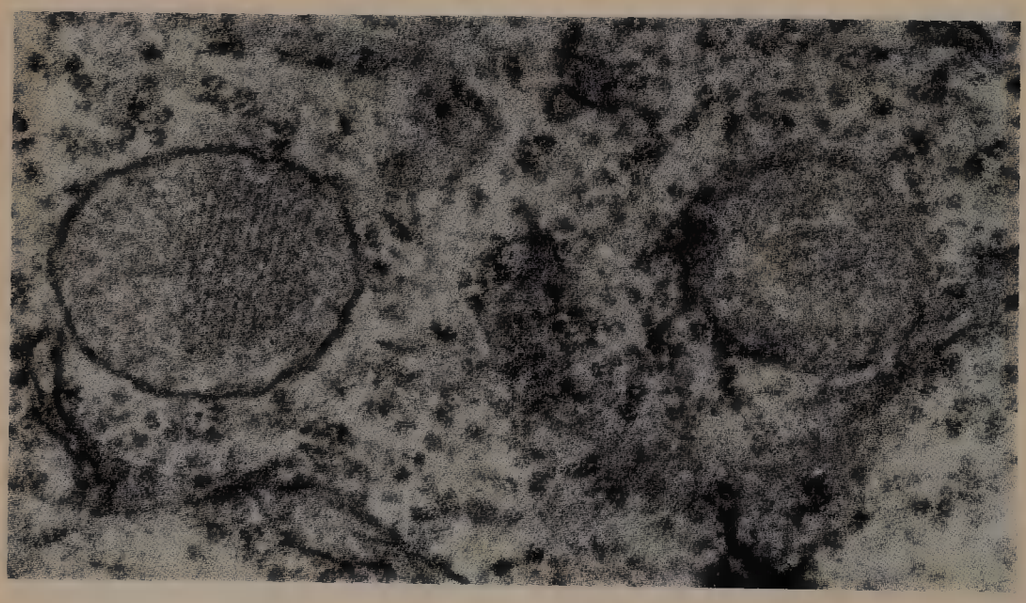
brane system. As I have indicated, much of what is known about the biogenesis of peroxisomes and glyoxysomes (there is no information on that of hydrogenosomes or glycosomes) does not support such an origin. What one sees in the case of peroxisomes—the posttranslational transfer of proteins through the cytosol—resembles instead the biogenetic pathway of many proteins that are encoded in the nucleus, synthesized in the cytoplasm and then introduced into mitochondria and chloroplasts. These are the two membrane-bounded organelles many biologists believe are derived from endosymbionts. The idea is that small bacterial cells, having respectively the ability to catalyze oxidative phosphorylation and to photosynthesize, were engulfed by a primitive ancestor of today's eukaryotic cells and stayed on to develop a symbiotic relation with the larger cells and eventually to become organelles. And so the question arises: Could microbodies too have originated as endosymbionts?

Now, the main evidence for such an origin in the case of mitochondria and chloroplasts is that both of those organelles contain their own DNA, specifying some of their own proteins, and the genetic machinery to replicate it, transcribe it into RNA and translate the RNA into protein. A few years ago there was a report that what looked like DNA could be seen in hydrogenosomes, but Müller has not been able to confirm the claim biochemically; there is no evidence that any of the microbodies has its own genetic system. A secondary argument for a symbiont origin of mitochondria and chloroplasts is that both of those organelles are bounded by a double membrane, the outer one of which may be a descendant of a vacuolar membrane that once housed the original bacterial symbiont. The microbodies (except for hydrogenosomes) are bounded by a single membrane.

These negative arguments do not make a watertight case against an endosymbiont origin for microbodies. One could point out that mitochondria and chloroplasts have lost to the nucleus more than 90 percent of their genetic information. Might the loss not be 100 percent in the case of organelles that may have been adopted as endosymbionts earlier? As for the two-membrane argument, not all endoparasites are enclosed in a vacuole. Some are naked in the cytoplasm.



HYDROGENOSOMES of *Trichomonas vaginalis* are enlarged some 40,000 diameters in this micrograph made by Shio. The large dense objects are rosettes of glycogen, a storage form of glucose. The striated object at the left in the micrograph is a riblike structure, the costa.



GLYCOSOMES, which are generally smaller than other microbodies, are enlarged 105,000 diameters in a micrograph made by Pierre Baudhuin. The particles are seen in *Trypanosoma brucei*, a protozoan parasite of domestic animals in Africa that is transmitted by the tsetse fly.

What makes the endosymbiont hypothesis attractive, besides the biogenetic argument already mentioned, is the primitive—one is tempted to say "bacterial"—character of the metabolic functions subserved by microbodies. The glycolytic chain is thought to be among the oldest enzyme systems, dating back to before photosynthesis, when there was virtually no oxygen in the atmosphere; it could survive in a particularly ancient form in the glycosome. The ability to reduce protons to molecular hydrogen, characteristic of the hydrogenosome, may also belong to early anaerobic days; it has been preserved in a similar form in *Clostridium*, a group of anaerobic bacteria that includes the agents of gas gangrene. As for the peroxisome, it could embody an early protective adaptation to the appearance of oxygen. A simple, nonphosphorylating peroxisomal type of respiration could have preceded the more advanced phosphorylating form found in some bacteria and in mitochondria. A final peroxisomal quality that is also characteristic of bacteria is ready adaptation to the availability of unusual nutrients.

These are flimsy arguments, to be sure. When it comes to events that happened more than a billion years ago, facts are scarce and imagination has an open field. The possibility that microbodies originated as symbionts would hardly be worth mentioning if it were not for the fact that it can be tested. One can determine the amino acid sequence of microbody proteins (or the nucleotide sequence of the genes encoding them) for comparison with the sequence of homologous molecules in various eukaryotic and bacterial cells, and the degree of relatedness may reveal something about the ancestry of microbodies. Such work is now in progress.

Museum piece, circa 1987.

The clapstick may be taking an early retirement. This and other conventional tools of the filmmaking trade may find themselves relegated to crates and carted off to museums.

What will replace them? Kodak film with Datakode magnetic control surface! This radical advance in film manufacture, which unites chemical and magnetic imaging technologies, gives film the ability to "converse" with computers. This achievement could substantially reduce the time and costs associated with film postproduction.

The Datakode magnetic control surface is a thin, transparent layer of iron oxide particles (approximately 9 billion per square inch) coated across the entire back of the film during manufacture. Less than 8 microns thick, this transparent layer provides the means to record machine-readable information (100 bits of binary data in a single track on each frame of film), and makes possible a uniform frame-indexing code that can be used with both film and videotape. Best of all, it accomplishes this interface without altering the quality or characteristics of the final image.

In the not-too-distant future, the use of discs, video displays, time-code synchronization, and automated printing equipment will speed filmmakers through all the noncreative, repetitive steps associated with film postproduction.

The new Datakode magnetic surface is more than just another milestone for Kodak and a boon to filmmakers. It represents the most significant development in motion picture technology in the last 50 years. It's a victory for science. And for the chemists and researchers who met the challenge to add to film's capabilities.

For more information on the new Datakode surface, send your request to: Eastman Kodak Company, Department GBSA-7, 343 State Street, Rochester, NY 14650.

Kodak

Kodak. Where technology anticipates need.

© Eastman Kodak Company, 1983

The new Honda Prelude began as a clean piece of paper.

Honda engineers sketched a superior aerodynamic shape. A low wedge with a wide stance and increased overall length.

Once the design was conceived, they set about perfecting it. They made the windshield and door handles flush with the body. New retractable halogen headlights raise in such a way to cause less turbulence.

Underneath is a new front suspension. It's a unique double wishbone, low-profile concept inherited from Honda Formula racing car experience. Independent MacPherson struts support the rear wheels. A low center of gravit contributes to handling and stability.

A new engine improves performance. It ha 12 main valves, 2 intake and 1 exhaust for each cylinder's main combustion chamber. The resu is better breathing and efficiency. And the du carburetors make highway passing power mo than passable.

Engine displacement is upped to 1829cc With the standard 5-speed shift, there's an increase of 33 percent in horsepower. There also a new 4-speed automatic transmission available. Variable-assist power steering come only with the automatic.

© 1983 American Honda Motor Co., Inc.

Particular attention was given to the new
erior. There's more room. Front bucket seats
ld you firmly and comfortably for extended
uring. Push-button controls are all within
sy reach of the driver. Your hands easily grip
e thicker 3-spoke steering wheel.

And when you aren't carrying passengers in
e rear seats, the rear seatback unlocks with
ey and folds down for added luggage space.
lso gives you access to the trunk. And trunk
e has been increased by 20 percent.

A power-operated Moonroof is included in
e long list of Honda standard equipment.
ong with our best AM/FM stereo radio with
digital electronic tuning and an autoreverse
cassette and four speakers.

When you first drive the new Prelude, you'll think you are in something else. You'll be right.

HONDA

The new Prelude.

SCIENCE AND THE CITIZEN

The Trouble with Trident

The decision by the Reagan Administration to accelerate the development of the Trident II sea-launched ballistic missile (SLBM) reflects among other things a growing appreciation of certain advantages inherent in deploying strategic nuclear weapons at sea rather than on land. Unlike land-based intercontinental ballistic missiles (ICBM's) such as the Minuteman and its proposed replacement, the MX, SLBM's on submerged submarines are difficult to locate accurately; hence they are essentially invulnerable to a preemptive "counterforce" attack and seem certain to remain so for the foreseeable future. Moreover, U.S. ballistic-missile submarines, by virtue of their range and mobility, are capable of launching a prompt retaliatory strike on the U.S.S.R. from many directions, thereby complicating any attempt to thwart a counterattack by means of an anti-ballistic-missile (ABM) system. In short, SLBM's come close to being an ideal deterrent force against a nuclear attack.

The mobility of ballistic-missile submarines, combined with the need for them to go undetected, has until now made SLBM's considerably less accurate than ICBM's and accordingly less suitable as counterforce weapons against "hard" targets such as missile silos and command, control and communications facilities. The Trident II, however, promises to overcome this limitation. Technological advances in missile-guidance systems are expected to give the Trident II an accuracy comparable to that projected for the MX. As a result the new generation of SLBM's should in principle be capable of destroying the entire range of potential targets in the U.S.S.R.

From the viewpoint of arms control the accelerated Trident II program must be considered a mixed blessing. On the one hand, it can be argued that greater reliance on submarine-based missiles would contribute to strategic stability; because submarines are apparently invulnerable, no attack could destroy the entire retaliatory force. Even if the Russians were to launch an attack, at least SLBM's would not draw fire to the continental U.S.

On the other hand, any analysis of counterforce-capable SLBM's ought to consider how the Russians are likely to react to the deployment of such missiles by the U.S. The U.S.S.R. now has most of its strategic nuclear weapons in the form of fixed land-based missiles. In addition it is generally agreed that the U.S.S.R. is far behind the U.S. in the development of an antisubmarine-warfare capability. Consequently Russian leaders might well perceive the Trident II program as a potentially dangerous first-strike threat to their ICBM force. One "quick fix" response could be the deployment of mobile land-based missiles. Another could be the adoption of a "launch on warning" policy, an outcome fraught with ominous possibilities, particularly given the reduced warning time available in the case of an actual or apparent SLBM attack.

The Administration's proposed military budgets for the fiscal years 1984 and 1985 allocate respectively $1.5 and $2.2 billion for development of the Trident II, up sharply from the current fiscal year ($369.7 million). If Congress approves the acceleration of the Trident II program, the first of the new missiles could become operational as early as 1988. As matters now stand, the full Trident II program calls for the deployment of 20 Trident submarines over the next decade, each one bearing 24 Trident II missiles; each missile in turn is reported to have a range of 6,000 miles and to be capable of carrying between 10 and 15 nuclear warheads with an explosive yield of 475 kilotons each. (The current generation of Trident I missiles are less accurate, have a range of 4,800 miles and carry between eight and 10 warheads of 100-kiloton yield.) The projected total cost of the Trident II program, including submarines, missiles and bases, is said to be roughly $60 billion.

Betting the Farm

The largest program of agricultural research in the U.S. is sponsored by the Department of Agriculture. Some of the work is done by the department's staff of scientific workers and some is done by outside agencies with funds provided by the department. Most of the in-house work is done by the Agricultural Research Service, which has a research budget of more than $400 million per year. In recent months the ARS has been the subject of considerable criticism both from within and from outside the Federal Government. Partly in response to the criticism, the agency has proposed a new research agenda that would increase the amount of work it does in basic science, particularly in areas related to new biological technology. The agenda would reduce the amount of work directed toward the improvement of plant and animal productivity, which has long been the largest single project carried out by the ARS.

Some of the most pointed criticism of the ARS came at a conference held last summer at Winrock, Ark. The conference was organized by Denis J. Prager, assistant director of the Office of Science and Technology Policy, and it included representatives of government, universities, foundations and private industry. *Science for Agriculture,* the report of the conference, concludes that the quality of research done by the Department of Agriculture has been declining steadily for several decades. According to the report, until the 1930's workers in the department did much of the fundamental work in basic science related to agriculture. Furthermore, because of its scientific preeminence, the staff of the department was able to provide strong formal and informal leadership in establishing priorities for all agricultural research done in the U.S.

In the past 50 years, however, the work of the ARS has become progressively more parochial, more fragmented and less fundamental in a scientific sense. Most projects are now aimed at solving specific local problems in animal and plant husbandry, such as how to eliminate a particular parasite in a small geographic area. As a result the department no longer exercises effective leadership in the setting of research priorities.

At least two outside influences have apparently contributed to a decline in the quality of ARS research. The first influence is that of Congress. Particularly since the 1950's, members of Congress, motivated by a desire to have Federal research facilities in their state, have been instrumental in the proliferation of ARS stations. There are now 148 ARS experimental stations in the U.S., ranging from the Agricultural Research Station at Beltsville, Md., with 450 scientific workers, to stations with only two or three workers. It is generally agreed that the cumbersome and inefficient institutional structure of the agency has contributed to its scientific decline, but attempts to close small facilities have met with much political resistance.

The second influence is a sharp decrease in the proportion of financial support for agricultural research done by state agencies that is provided by the Federal Government. Under the Hatch Act of 1887 the Government distributes funds for agricultural research to state governments. The amount depends on the size of the state's rural and farm population. Most of the money goes to support the state agricultural experiment stations. The states are required to match the amount provided by the Department of Agriculture, which is now about $140 million. In recent years, however, the contributions of the states have far exceeded those of the Depart-

We're keeping the feathers dry.

Before the last four flights of America's most famous bird, the Columbia space shuttle, our "Scotchgard" Brand Fabric Protector was sprayed right from the can onto the shuttle's surface. It's the same product you buy to protect clothes and furniture.

On the Columbia it kept rain from seeping into the surface tiles.

That solved a problem from the Columbia's first flight, when moisture in the tiles froze in the cold of space, expanded, cracked the tiles and endangered a safe re-entry.

When people ask us to solve tough problems, we often have the solution right on the shelf.

Then too, by listening to people's needs, we've pioneered over 200 high-performance consumer products. We make everything from sandpaper and skincare products to home video and computer tape.

And it all began by listening.

3M hears you...

For your free 3M Consumer Products Brochure, write: Department 070705/3M, P.O. Box 4039, St. Paul, MN 55104.

Name_____

Address_____

City_____ State_____ Zip_____

Or call toll-free: **1-800-323-1718,** Operator 367. (Illinois residents call 1-800-942-8881.)

3M

ment of Agriculture: their contribution is now about $700 million.

The effect of both congressional and state influence has been to shift research in the direction of applying well-known technologies and scientific results to the solution of local agricultural problems. Obtaining new fundamental scientific results and developing new technologies has largely been left to universities and private industry. The new ARS research agenda, which covers the years from 1984 through 1990, represents an attempt to reverse these trends. The proposal assumes that the current ARS research budget of $413 million will remain unchanged through 1990, after the effects of inflation have been taken into account. Under the proposal no major work now done by the ARS will be eliminated, but funds will be redistributed among the existing categories of research. Work on the productivity of crop plants and farm animals, which currently constitutes 60 percent of the ARS budget, would decrease to 52 percent by 1990. Most of the savings would be distributed among projects concerned with human nutrition, soil and water conservation, and commodity conversion and delivery.

Funds would also be redistributed within the major categories. For example, under the heading of plant productivity, work on insects, diseases and nematodes in field horticulture would decrease from $27 to $14 million. In the same category work on maintaining the genetic diversity of crop plants would increase from $10 to $12 million. A total of $75 million would be redistributed among existing projects.

Among those who attended the Winrock conference there are sharp differences of opinion on the new direction at the ARS. John A. Pino, director for agricultural sciences of the Rockefeller Foundation, described the agenda as "a move in the right direction [that] provides a stronger focus on some of the basic aspects of agricultural science." Pino added that in his opinion the ARS had a responsibility to do work that would lead to "an understanding of some of the fundamental processes... involved in the interaction of plants and animals." He said of the current proposal: "I hope it's more than a cosmetic change." He noted, however, that the agenda merely redistributes funds without eliminating any of the work currently being done. To make a fundamental change in the direction of research, substantial cuts will have to be made, he said. "Some surgery is going to be needed. It remains to be seen whether the surgery can be done and upheld."

James T. Bonnen, professor of agricultural economics at Michigan State University, on the other hand, questioned the notion that the solution to the problems of the ARS is an increased commitment to basic science. According to Bonnen, "some critics of the agency talk as if the ARS ought to do only basic science. If that is so, how do you get the very large national agenda of applied work done? If Federal agencies won't do it, it won't get done." The primary role of the ARS, he said, should be to work on national problems in applied science and technology. "The people who are critical of the ARS come from one end of the spectrum: basic science. They apparently have no appreciation for the intervening institutions between basic science and [agricultural] productivity."

Living Color

In recent years the most important tool for the study of how the brain sees has been the microelectrode. Inserted into the brain of experimental animals, it has revealed that nerve cells in the retina, in the cerebral cortex and in the lateral geniculate nucleus (a group of cells interposed in the visual pathway that leads from the retina to the cerebral cortex) respond preferentially to particular patterns of lightness and darkness in the visual field. The neural mechanism underlying the perception of color, however, has remained elusive. Two recent discoveries promise to change that. One of them was made with microelectrodes in the brain of the monkey; the other is based on investigations of how colors look to a human observer under special circumstances. The new discoveries support the growing consensus that the brain performs several simultaneous analyses of the visual world, processing visual information in a number of more or less independent ways before reintegrating the data to produce the unified conscious sensation of sight.

Some of the simultaneous analyses the brain brings to bear on vision became apparent in the 1950's and 1960's as David H. Hubel and Torsten N. Wiesel of the Harvard Medical School described the physiological properties of nerve cells in the visual cortex, the specific part of the cerebral cortex receiving visual data from the retina by way of the lateral geniculate nucleus. They worked mostly with cats. The cells with the simplest properties were found to be most responsive (that is, they changed their pattern of electrical activity most markedly) when the investigators presented a small bar of light at a particular orientation in a particular part of the cat's visual field. If the orientation was altered or the bar was enlarged, the cell became less responsive. Indeed, its electrical activity could be inhibited. Evidently the cell reacted to "on" regions and "off" regions in the visual field. In many cases the regions were arrayed so that the resulting "receptive field" of the cell (the locus in the visual field to which the cell is sensitive) had an "on" center and an "off" surround.

Other cells in the visual cortex have more complex receptive fields. They respond best to discontinuous bars of light at particular places and orientations. An overall pattern emerged, in which nerve fibers carrying visual data from the lateral geniculate nucleus pass their signals to cells in a certain layer of the visual cortex; it is designated layer 4*c*. The cells in this layer monitor specific parts of the visual field but have no specificity for the orientation of the stimulus. They communicate, however, with cortical cells in layers 2 and 3, and the communications are "wired" in ways that make the latter cells quite sensitive to stimulus orientation. The cells in layers 2 and 3 communicate in turn with cells in layers 5 and 6 in ways that give many of the cells in these deepest cortical layers still more complex sensitivities. The pattern offered tantalizing hints of how the brain might analyze images for their forms, or at least for features such as edges and corners. It offered no hint of how the brain sees colors.

While Hubel and Wiesel were making their discoveries, Edwin H. Land and his colleagues at the Polaroid Corporation were formulating the retinex theory of color vision. Land began by noting that the colors of a scene look much the same to an observer no matter how the scene is illuminated (in sunlight, for example, and then in deep shade). Moreover, a scene illuminated with artificial light in a quite narrow range of wavelengths often turned out to retain a surprising wealth of perceived colors. Land concluded that the colors seen by an observer have no simple relation to the wavelengths of the light entering the eyes. The colors result instead from computations performed by the brain.

The retina is known to have three distinct types of cone cells, which are the light-receptor cells sensitive to color. One type is most sensitive to the shorter wavelengths of visible light, the second to the middle wavelengths and the third to the longer wavelengths. Land therefore proposed that each type of cone cell supplies information enabling the brain to compute ratios of brightness in the corresponding range of wavelengths. The ratios would be calculated at the boundaries between objects of differing color encountered along arbitrary paths across the visual field. The process amounts to the assignment of three brightness values (short-, middle- and long-wavelength "lightness") to any given part of the visual field. The triplets of brightness values are scaled by conditions that apply across the entire field; thus the ratios are not much affected by how the field happens to be illuminated. As a result the triplets can yield more or less invariant color sensations.

"Since we are uncertain of the loca-

Introducing the new Osborne™ Executive™ Series.

We listened when you told us you wanted a personal business computer.

We delivered the Osborne 1.

And in the two years since its introduction, the Osborne 1 has become the largest selling portable business computer on earth.

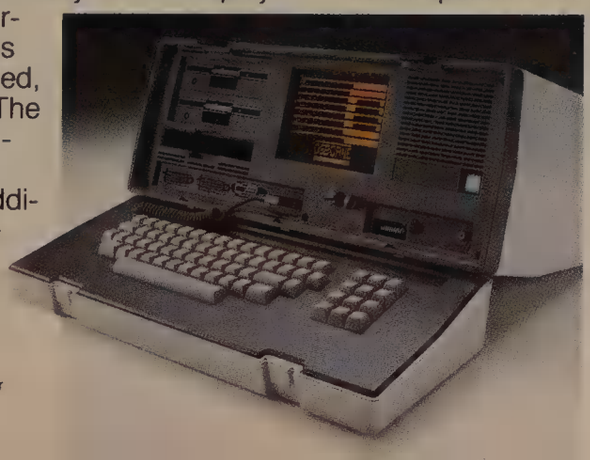
We listened harder.

The result is the new Osborne Executive Series. Consider these new advantages:

☐ A big, 80-column, eye-easy amber display lets you see more of the information you're using, with less effort. ☐ Memory is expanded, for larger spreadsheets. ☐ The new Executive Series now includes a database system—Personal Pearl.™ (That's in addition to an already formidable list of standard Osborne software, including word processing and electronic spreadsheets.) ☐ Terminal Emulation provides access to information stored on other computers. ☐ IBM Personal Computer™ compatibility (on the Executive 2) lets you use software packages designed for the IBM system.

And there's much more.

Add up the value of all the hardware features and software systems included with either the Executive or Executive 2 and you will see that the sticker price is one of the incredible bargains available in personal computing.



Visit your local authorized Osborne dealer. Once again, Osborne has set the standard for portable business computers.

For the name of your nearest Osborne dealer, dial (800) 772-3545 ext. 905 (in California); (800) 227-1617 ext. 905 (outside California).

OSBORNE™
COMPUTER CORPORATION

Trademarks: Osborne 1, Executive, Executive 2: Osborne Computer Corporation. Personal Pearl: Relational Software. IBM Personal Computer: International Business Machines.

tion of the mechanisms that mediate these processes," Land wrote in 1977, "I have coined the term retinex (a combination of retina and cortex) to describe [them]." Land, in collaboration with Hubel and with Margaret Livingstone of the Harvard Medical School, now reports the results of an experiment that enables the investigators to specify a location in the brain: it is the cerebral cortex. In each trial of the experiment subjects were asked to call out a succession of small letters flashed rapidly at the same place (a "fixation point") on a screen. Well to the left of the fixation point lay a collage of colored paper. Attending to the flashing letters ensured that the collage lay entirely on the left side of each subject's visual field.

The collage could be illuminated by light shone through various filters. Each filter, however, had a hole cut in it, so that the illumination of a small test spot surrounding the fixation point never changed. If the perception of color depended on only the wavelengths of light, the test spot would always look the same. In fact it did not. "With the filters in place," Land and his colleagues write in a paper to be published in *Nature,* "the test spot was a chalky white, whereas with no filters it was a deep purple—although the flux from the test spot had not changed at all." In accord with the retinex theory, the altered illumination of the collage had affected the perception of color throughout the visual field.

Next Land and his colleagues enlisted the assistance of a 29-year-old man, identified as J.W., who had undergone neurosurgery two years earlier in a successful effort to reduce the number and severity of his epileptic seizures. The surgeon had cut the corpus callosum: a massive plate of nerve fibers that connect the cortex of the left and right cerebral hemispheres. Like other split-brain patients, J.W. showed no obvious subsequent impairment. His color vision was normal. Tests determined that the parts of the cerebral cortex specialized for language were on the left side of his brain. Thus J.W. could give verbal descriptions of things in the right half of his visual field. (Information from the right half of the field passes from the eyes to the left side of the brain.) Things in the left half of the field could not be described; information from the left visual field continued to pass from the eyes to the right side of the brain, but since the corpus callosum had been severed, the brain no longer had a path by which information could cross from the right side to the language centers on the left.

When J.W. was tested with Land's apparatus, the results were remarkable. At first the collage was in the left half of the visual field. The test spot spanned the vertical midline, which divides the halves of the field. J.W. "consistently reported that the test spot was 'white' regardless of the illumination." Then a mirror was interposed, so that the collage was in the right half of the visual field. J.W.'s first comment was that "he saw 'all colors....' It turned out that he was describing the entire [collage], something he had apparently not been able to do when it was presented to his nonverbal hemisphere." After that he described the test spot as purple or as white, "in complete agreement with the appearances to a normal observer."

If the retinex calculations were done by neurons in the retina, whose grasp of the visual world extends well across the vertical midline, the severing of the corpus callosum would have no effect on the perception of color. Each side of the brain would get the same color sensations based on identical retinex calculations, and J.W.'s responses would always duplicate those of a normal observer. J.W.'s responses did not. They show that the retinex calculations need the corpus callosum to provide a path from one side of the brain to the other. "The calculations," Land and his colleagues conclude, "must [occur] in the cerebral cortex."

Without the second recent discovery

If you drink water, you know our work. Now we'd like you to know our name.

Experienced in every phase of water resource development, URS is one of the leading international organizations providing engineering, architectural planning, environmental and construction management capabilities within one company. Our professional staff provides an integrated approach to large and complex assignments. Learn more about our many capabilities and how we can help you.

URS CORPORATION

Phone or write: URS Corporation, 155 Bovet Road, San Mateo, CA 94402 (415) 574-5000

Making Technology Work™

Land's finding would have been vexing, because the cells in the cerebral cortex known to be sensitive to color are incapable of performing retinex calculations. Opponent-color cells, for example, had been known since the late 1950's. Gunnar Svaetichin of the Venezuelan Institute of Neurology and Brain Research and Edward F. MacNichol, Jr., of Johns Hopkins University first described them in the retina of various shallow-water fishes, and they have since been found by Hubel and Wiesel in the visual cortex of the monkey. Like other cells in the visual cortex, they respond in opposite ways to "on" and "off" regions in their receptive field. In the case of the opponent-color cells, however, the response depends on the wavelength of light. An opponent-color cell might be excited by a small red spot and inhibited by a large field of green. ("Red" and "green" refer here to wavelengths of light, not to perceived colors.) Presumably the input to the cell derives ultimately from at least two types of cone cells. The problem is that a red spot on a green background would leave the cell unaffected. The excitation and the inhibition would cancel each other. Since the cell could make no response to a boundary between contrasting colors, it could make no contribution to retinex calculations.

What is needed for the calculations is a network based on what are termed double-opponent-color cells. A cell of this type might be excited by, say, a small red spot in a particular part of the visual field. It would be inhibited by a small green spot in the same position. The colors surrounding the position would have the opposite effect: red would be inhibitory and green would be excitatory. The cell would thus make no response to white light of any pattern, nor to diffuse light of any wavelength. It would make no response to a large expanse of red or of green; the center excitation and the surround inhibition would cancel. Its greatest excitation would be in response to a red spot on a green background, its greatest inhibition in response to a green spot on a red background. Hence it is specialized to respond most markedly to contrasting colors.

In the 1960's Nigel W. Daw, who is now at the Washington University School of Medicine in St. Louis, found double-opponent-color cells in the retina of the goldfish. Soon a number of investigators began to find them in the visual cortex of the monkey. In all accounts, however, the cells were disconcertingly few and gave no sign of being concentrated in any particular cortical locus. Then in 1978 Margaret T. T. Wong-Riley, who is now at the Medical College of Wisconsin, stained the visual cortex of the monkey by a technique that reveals the distribution of cytochrome oxidase, an enzyme found in mitochondria. Mitochondria are the organelles that produce energy in living cells; thus the stain is thought to mark sites of great activity in a tissue. Wong-Riley found dense concentrations of cytochrome oxidase in flask-shaped blobs about 200 micrometers in width spaced at .5-millimeter intervals across the visual cortex. In 1980 it emerged that the blobs form a polka-dot array in layers 2 and 3. It has since emerged that the blobs receive nerve fibers from the lateral geniculate nucleus.

Hubel and Livingstone employed a microelectrode to probe the cells in the blobs. None of them was sensitive to the orientation of a bar of light. Apparently they take no part in the analysis of edges and corners. On the other hand, as many as 90 percent of them are sensitive to color. Most are double-opponent-color cells.

Vaccine Gene

Two of the powerful new tools of molecular biology, monoclonal antibodies and the molecular cloning of recombinant DNA, have been exploited

Unmistakably Zeiss

When a binocular carries the Zeiss name, it's your assurance of superb optical quality and unerring mechanical precision. Zeiss binoculars offer outstanding contrast, color definition and clarity. And no one surpasses Zeiss in light transmission, which means brilliant images at dawn and dusk.

Zeiss binoculars...recognized the world over for the highest quality and performance. And backed by a lifetime service guarantee.

For the Zeiss dealer nearest you, call toll-free (800) 446-1807 and ask for Department 207.

Carl Zeiss, Inc.
Consumer Products Division
1015 Commerce Street
Petersburg, Virginia 23803

ZEISS
West Germany

The great name in optics

THE TECHNOLOGY OF THE CENTURY.

THE 1983 BUICK CENTURY.

With today's technology, almost anything is possible. Computers that talk. Computers that teach. Electronics that put Wall Street right in your living room. And the Buick Century. A car so technologically advanced, it makes driving a pleasure instead of a chore.

The way Century accomplishes this is quite remarkable. There's an available electronic digital instrument panel, for example, that's a lot more than needles and dials. It tells you mph, km/h, fuel level, even how far you've traveled. All in computerized digital display.

Century's standard engine is electronically fuel injected. And connected to front-wheel drive. Its ride and handling are computer tuned. And all of this is surrounded by an aerodynamic shape that was refined in the wind tunnel, and further refined by computer.

Then there are, of course, the little things. Like an electronic tone for the ignition key and seat belt, instead of a buzzer. Why, Century is even built with robotic assembly techniques.

Happily, there remains one human quality that technology, even the technology of the Century, can never replace. The sheer pleasure of driving a Buick. And that's something you can experience at your Buick dealer now.

EST. HWY. 39	EPA EST. MPG 24

Use estimated MPG for comparison. Your mileage may differ depending on speed, distance, weather. Actual highway mileage lower. Some Buicks are equipped with engines produced by other GM divisions, subsidiaries, or affiliated companies worldwide. See your Buick dealer for details.

Official Car of XXIIIrd Olympiad Los Angeles 1984

© 1980 L.A. Olympic Committee ™

BUICK

Wouldn't you really rather have a Buick?

to make the development of a vaccine against malaria feasible in principle. The manufacture and testing of such a vaccine is probably some years away, however, in part because of a disagreement between the biotechnology company that would manufacture the vaccine and an international agency that has helped to support the research.

After World War II there was reason to hope that a combined attack on *Anopheles* mosquitoes with DDT and on *Plasmodium* parasites with new specific drugs might soon bring malaria under control. For a time there was progress. Now the incidence of malaria is on the rise, however, and so is the death rate. The logistics of spraying and of drug delivery have been daunting in underdeveloped regions and particularly in tropical Africa, where malaria is endemic and a million children die of the disease every year.

In this situation a malaria vaccine would be an ideal weapon, but the plasmodium is a difficult target for immunologists. Four species cause malaria in human beings and each one carries on its outer surface a different set of antigens, the proteins that provoke (and become the targets of) an immune response by antibody-secreting cells. The antigens may also vary in different strains of a single species. Indeed, the plasmodium displays different sets of antigens successively in the several morphological stages of its complex life cycle in the mosquito and in the human victim. The mosquito's bite injects sporozoites, which infect liver cells and divide to form small merozoites. Each merozoite invades a red blood cell, grows through several stages and divides into from 12 to 24 new merozoites that burst out of the cell and invade other cells. A few merozoites develop into gametocytes, whose ingestion by a biting mosquito initiates a new cycle and spreads the disease.

The complexity of the various parasite forms and their antigenic variability have made it clear that intact parasites, whether attenuated or killed, will not serve as the active agent of a vaccine. The effort recently has been to develop a vaccine by finding and isolating a particular antigen, preferably one whose role in infection is so important that it precludes much variability from strain to strain. For some years Ruth and Victor Nussenzweig of the New York University School of Medicine and various collaborators have taken that approach. They have isolated from the surface of the sporozoites that cause a number of animal and human malarias a protein whose composition is similar in all species and that has a role in the process of infection. By crossing specific antibody-producing cells with cells of the tumor called a myeloma they have made hybridoma cells that secrete monoclonal antibodies directed specifically against the sporozoite surface protein. Injected into chimpanzees, the antibodies give a measure of passive immunity against malaria, suggesting that the surface protein should serve as an effective antigen.

To make the large quantities of the protein needed even for testing a vaccine one might dissect millions of mosquito salivary glands and isolate the sporozoites and then the protein. An alternative is to find the gene encoding the protein, introduce it into *Escherichia coli* and get the bacteria to manufacture the protein. That is what G. Nigel Godson of N.Y.U., with his students Joan Ellis and Luiz Ozaki, has done.

Godson's method, which is reported in detail in *Nature*, relied heavily on the existence of the Nussenzweigs' monoclonal antibodies. He extracted messenger RNA (the nucleic acid into which the DNA of genes is transcribed and that in turn is translated to make protein) from mosquitoes infected with *P. knowlesi*, a plasmodium that causes malaria in monkeys. He made DNA from the RNA, broke up the DNA into thousands of short fragments and inserted each fragment into a plasmid: a small circular piece of bacterial DNA that serves as a vehicle for cloning. Each plasmid was inserted into bacteria. To find any bacteria that were synthesizing the *P. knowlesi* surface protein, and must therefore carry on a plasmid the gene encoding the protein, he tested each bacterial colony with the monoclonal antibody to the protein. When he found such bacteria, he grew their plasmid in quantity and showed by various tests that it did indeed carry the gene for the designated protein. The next step will be to show that when the protein made in bacteria is injected into monkeys, it confers immunity to malaria.

With the gene for the *P. knowlesi* protein in hand, it should not be hard to find the gene (which must be quite similar) for the analogous protein of *P. falciparum*, the agent of the lethal form of human malaria. That effort is under way at N.Y.U. Meanwhile the institution has filed for a patent on the process and has been negotiating an agreement licensing Genentech, Inc., a San Francisco company that specializes in recombinant-DNA technology, to develop, test, manufacture and market a vaccine. Any income to N.Y.U. would be devoted to research in tropical medicine. The licensing agreement requires the approval of public agencies that have helped to fund the work.

One of the agencies, the World Health Organization (WHO), objected to a provision giving Genentech exclusive marketing rights, without which the company is reluctant to develop the product. WHO is presumably worried less by the financial aspects of the agreement than by the possibility that Genentech, having tied up rights to the vaccine, might never market it. Agency officials have often complained that drugs having critical importance for the underdeveloped world are sometimes not made available because there is no profitable market for them in affluent countries.

Prime Time

When a computer that can do a million operations per second is responding to keystrokes made at intervals of some tenths of a second, it can have a lot of spare time. Paul A. Pritchard of Cornell University decided to put a computer to work in those idle moments. He instructed the computer, whenever it was not busy doing something else, to search for a series of at least 18 prime numbers that advance in an arithmetic progression. (A prime number is divisible only by itself and 1; in an arithmetic progression the difference between each number and the next is the same).

Pritchard wanted 18 such prime numbers because the previous record, set in 1977, was 17. Nipping in for a microsecond here and a millisecond there on a computer that is devoted mainly to word processing and runs 24 hours per day, he was able to find a series of 18 arithmetic-progression primes in less than a month. The computer was actually working on the problem for an average of about 10 hours per day. The series begins with 107,928,278,317 and rises by an increment of 9,922,782,870 to 276,615,587,107. The numbers before the first in the series and after the last (taking into account the same increment) are not primes.

The strategy for the search was based on the fact that in any arithmetic progression of n primes the common difference must be divisible by all the primes less than or equal to n, except in the rare case of a series whose first number is n. In Pritchard's search n was 18, and so the common difference had to be divisible by 2, 3, 5, 7, 11, 13 and 17. Pritchard also included 19 in case the computer hit on a series of more than 18 arithmetic-progression primes. The common difference must be a multiple of the product of these eight primes, namely 9,699,690. Pritchard therefore had the computer choose a prime in some designated range of numbers, construct an arithmetic series by repeatedly adding a multiple of the product and test the resulting numbers for primality.

Pritchard had his in-and-out program going on two computers, each a fairly standard machine that is not notably fast. One of them is still searching after thousands of hours. Pritchard is continuing the search with both machines. He expects to find more 18-term primes in arithmetic progression and hopes to find one with 19 terms.

Modern Pork Production

Pork currently provides some 25 percent of the energy and 9 percent of the protein human beings get from animal sources. The interaction of economics and biology has given rise to a new kind of pig farming

by Wilson G. Pond

The pig lingers in the memory of many people as a farmstead animal that roots in the ground for various edibles, wallows in mud and is fed slop and swill, meaning the liquid and semiliquid leavings from the farmer's kitchen and the operation of the farm. That is not the pig of modern agribusiness. Such a pig lives indoors for its entire brief life: born and suckled in a farrowing unit and raised to slaughter weight in a nursery and later in a growing-feeding unit. It is fed a computer-formulated diet based on cornmeal and soybean meal with supplements of protein, minerals and vitamins. Unless it is destined for breeding, it is sent to market at five or six months of age, having reached the slaughter weight of 220 pounds or more from its birth weight of about two pounds. In the U.S., which is one of the leading countries in the production of pork, some 97 million pigs with a value of $8.9 billion went to market in 1980.

It is best to be precise about terminology. All pigs, wild and domestic, are swine, members of the family Suidae, which in turn is a member of the mammalian order Artiodactyla (even-toed ungulates). In the U.K. all domestic swine are called pigs, but in the U.S. the term applies technically to young swine weighing less than about 120 pounds; the others are called hogs. One hears also of boars (uncastrated males), barrows (castrated males), gilts (virgin females), sows (females that have bred) and shoats (newly weaned piglets).

The domestic pigs of today are descended from the European wild pig (*Sus scrofa*), which is also known as the European wild boar. Pigs living in the wild are still found in many regions of the world. Except for old males, which are solitary, wild pigs live in groups. They are nocturnal and omnivorous, digging for edible roots and tubers, gathering fruits and nuts on the ground and also eating a large variety of insects, reptiles, birds and small mammals. The wild pig stands about 90 centimeters (35 inches) tall at the shoulder and has a thick coat of bristly hair. It is a fast runner and a good swimmer.

The time of domestication of the pig is lost in history. One study has it that the earliest known domestication of pigs was in what is now Iraq in about 6750 B.C. A Chinese scholar has asserted that domesticated pigs were to be found in his country by 2900 B.C. Probably the attractiveness of the pig as a domestic animal originally was that it is an efficient scavenger and will eat a wide variety of foods. That is still its role in many developing countries.

In modern agribusiness the pig is attractive because of its efficiency in converting feed into food. Today's pig gains a pound of weight for each three or 3.5 pounds of food it eats. Fifty years ago it took more than four pounds of food to achieve the same result. In terms of the pig's yield of edible energy per calorie consumed it outdoes cattle, lambs and poultry. Indeed, the production of pork, like that of beef, lamb and poultry, is based on the economic advantage for the farmer of selling his crops through animals at a higher profit than he would realize by selling the crops directly for human food.

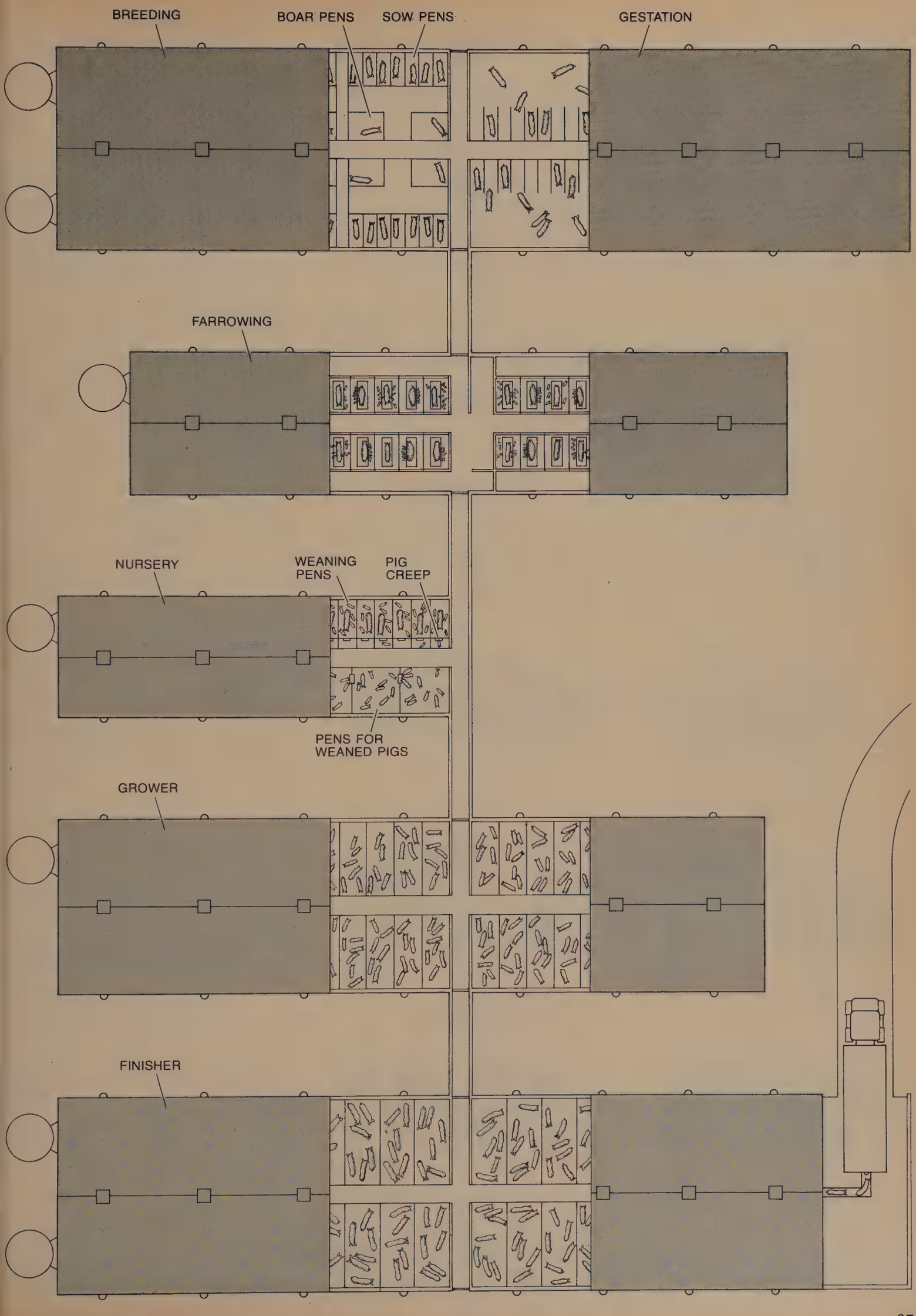
From the consumer's point of view the meat of the pig has high nutritional values. A four-ounce serving of lean pork supplies about three-fourths of the adult's daily need for thiamin and iron, from a fourth to a third of the need for niacin, vitamin B-6 and vitamin B-12, half of the need for protein and most of the requirement for trace minerals. In North America and Europe the annual consumption of pork is about 60 pounds per capita; the worldwide average, however, is less than 20 pounds per capita. Since the demand for animal protein in the diet increases linearly with average per capita income in a country, the geographic disparities in the consumption of pork largely reflect the huge disparity in per capita income among countries.

Modern Practices

The trend that took the pig out of the pigpen and into a carefully controlled environment resulted from high labor costs, the rise in land values, the fact that it became more profitable to raise crops on productive land than to keep animals on it, the ability to keep the pig's diet under closer control indoors and the possibility for more effective control of diseases and parasites. On a farm that raises pigs from birth to marketing age, which is a strong modern trend, a typical physical arrangement includes a unit housing pregnant sows, a farrowing unit where the sows give birth, a nursery and a growing-finishing unit. They all have slatted floors for easier removal of wastes. In Temperate Zone farms each unit (except sometimes the growing-finishing unit) has a heating system.

The main task of husbandry is feeding the pigs, an operation that accounts for from 55 to 85 percent of the total cost of the commercial production of pork. The variations result from the relative costs of feed, labor and housing between one farm and another or from one season to the next. The economics of feeding swine depend on the local availability and price of feedstuff, competition for a particular foodstuff for consumption by people or by other domestic animals and the prices of the protein, mineral and vitamin supplements.

TOTAL-CONFINEMENT UNIT depicted on the opposite page keeps pigs indoors from birth until they are sent to market. The perspective is straight down on a five-building complex, with sections of the roofs removed to show the activity inside. In the farrowing unit pigs are born and suckled while the sow remains in a metal crate that prevents her from rolling onto her piglets. In the nursery, the grower and the finisher the pigs are fed carefully formulated diets designed to bring them to the slaughter weight of 220 pounds or more at the age of five or six months. Many farms that raise pigs in this way have the grower and finisher as one building.


BREEDING
BOAR PENS
SOW PENS
GESTATION
FARROWING
NURSERY
WEANING PENS
PIG CREEP
PENS FOR WEANED PIGS
GROWER
FINISHER

The pig's digestive system and the human one are similar, so that the pig farmer is in direct competition with the human population for food supplies. The degree of competition is related to cultural differences in food preference. For example, in the U.S. wheat and potatoes are usually not fed to pigs because the human demand for these crops holds the price too high, whereas in other parts of the world pigs often get wheat and potatoes in their diet.

The farmer must plan on using up about 800 pounds of a typical concentrated diet based on cornmeal and soybean meal for each pig marketed at 220 pounds. The total includes feed that goes to the breeding herd and also takes into account the fact that even in a well-run operation a fair number of pigs die before reaching marketable age. Of the 800 pounds of feed from 100 to 200 pounds consist of the protein, mineral and vitamin supplements needed to make up the nutritional deficits in corn and cereal grain. The nutritional requirements of pigs at various stages of production have been so precisely determined that commercial feed manufacturers now formulate diets by computer. The number of farmers working the diets out on their own computers is increasing.

Although unprocessed corn and cereal grains make up a major part of most diets for pigs, the computer-formulated least-cost rations contain progressively higher proportions of by-product foodstuffs, including wheat and corn bran from mills, grains and solubles from distilleries and blood meal, meat meal and bone meal from slaughterhouses. Such by-products would have little value if they were not sold as feed for livestock; if they were simply discarded, the cost of disposing of them would significantly increase the market price of the final product.

It has also been found that forage crops such as alfalfa can be incorporated in the diet of pigs, particularly pregnant females. The reproductive performance of such females has been shown in recent tests to be normal when high-quality alfalfa meal constitutes 96 percent of the diet, with mineral and vitamin supplements making up the other 4 percent.

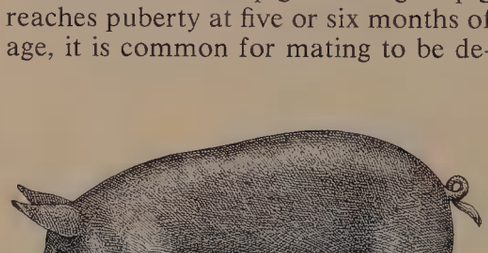
Biology of the Pig

The yearly production calendar of the pig farm must be geared to the biological timetable of the pig. Although a pig reaches puberty at five or six months of age, it is common for mating to be delayed until about eight months so that the animal will be larger and physiologically more mature. The gestation period is about 114 days, and the lactation period of the female continues for from six to eight weeks. The estrous cycle takes up about 21 days and repeats throughout the year. Pregnancy halts the cycle, which is also inhibited in the pig (as it is not in the cow) by lactation. The cycle resumes in from three to 10 days after weaning, which normally is done when the pigs are from four to six weeks old.

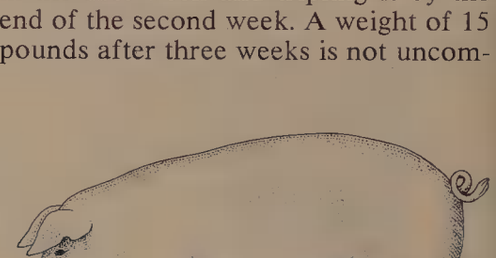
The pig's natural course of development dictates that the female can complete her first pregnancy at about one year of age and that she can complete no more than 2.7 cycles of reproduction and lactation per year thereafter. In practice the average reproductive rate of a herd is closer to 1.8 litters per year after adjustments are made for barren sows, failure to conceive on the first mating and other factors. A litter consists typically of nine or 10 piglets, of which 7.2 on the average survive to maturity.

Although a newborn pig weighs only about 2.2 pounds, it grows at a phenomenal rate, normally doubling its weight in the first week and tripling it by the end of the second week. A weight of 15 pounds after three weeks is not uncom-

SUS SCROFA



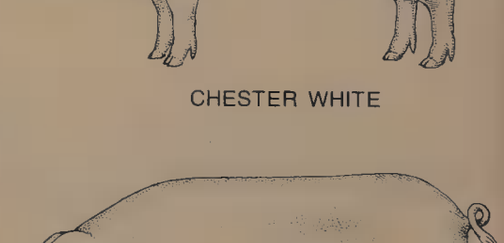
BERKSHIRE



CHESTER WHITE

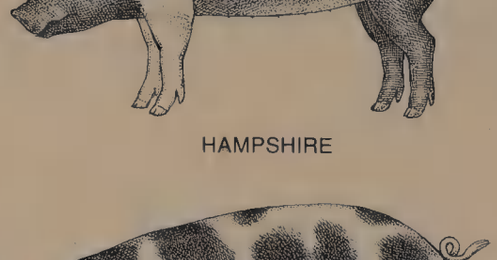
DUROC

HAMPSHIRE



LANDRACE

POLAND CHINA



SPOT

YORKSHIRE

PUREBRED BREEDS provide the stock for the crossbred herds found on most commercial pig farms. Also portrayed is the ancestral pig, the European wild boar (*Sus scrofa*). The Berkshire and the other domesticated purebreds shown here constitute the seed stock for most of the crossbreeding in the U.S. A few countries raise mainly purebreds, such as the Landrace herds predominating in Denmark.

mon. A market weight of from 220 to 240 pounds is attained by about five months of age, but individual pigs are known to have reached 200 pounds in less than 120 days having eaten less than two pounds of feed per pound of gain in weight.

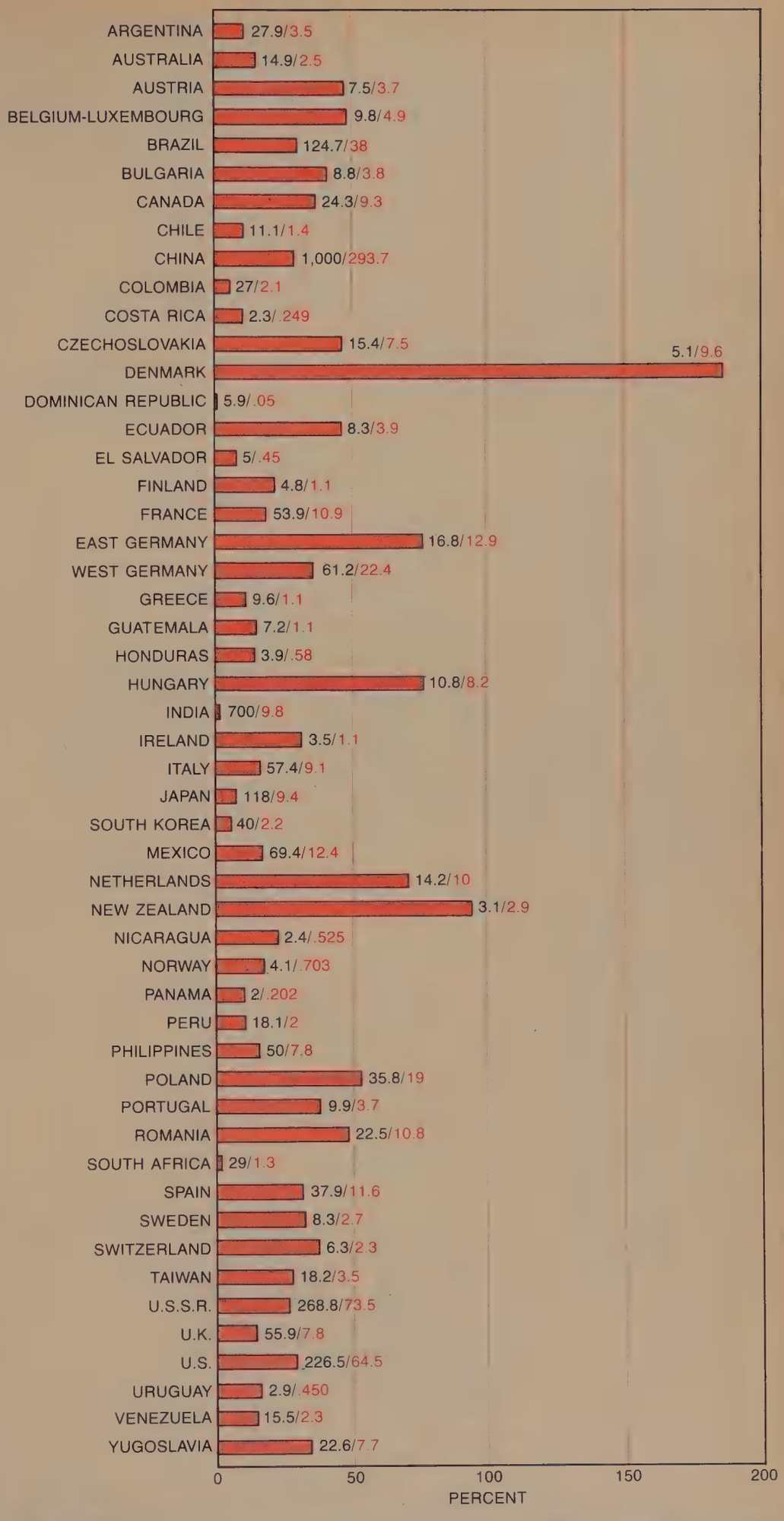
Male pigs are normally castrated during the first few days or weeks of life to avoid the marketing of pork with the objectionable taste associated with the meat of mature boars. The surgery can be done much later, however, with the same effect, and it is noteworthy that intact males grow faster and are leaner than castrates. As a result research is under way in several countries to develop mature boars that yield meat lacking the objectionable taste. Farmers often delay the castration of potential breeding boars in order to have more time to choose the ones that seem likely to make the best sires.

Role of Crossbreeding

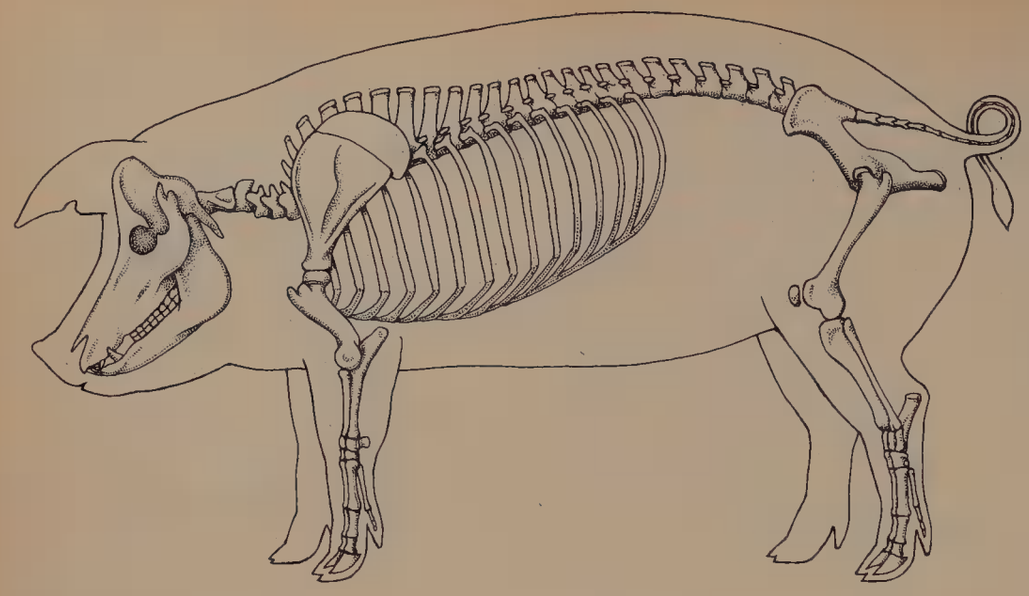
The genetic base for improving the breed in economically important ways is a group of purebred populations that constitute less than 10 percent of the total population of swine in most countries. There are exceptions, as in Denmark, where most of the pigs produced are purebred Landrace. In most other developed countries a few major breeds provide the seed stock for a commercial pork-production industry based on crossbreeding. In the U.S., for example, some eight breeds fulfill that role. In addition to the Landrace they include the Berkshire, the Chester White, the Duroc, the Hampshire, the Poland China, the Spot and the Yorkshire. Most of the swine breeds that figure in commercial production throughout the world originated in the U.S. or the U.K. The Yorkshire of the U.S. is similar to the Large White of Europe, and the American Landrace is similar to the Landrace of the Scandinavian countries.

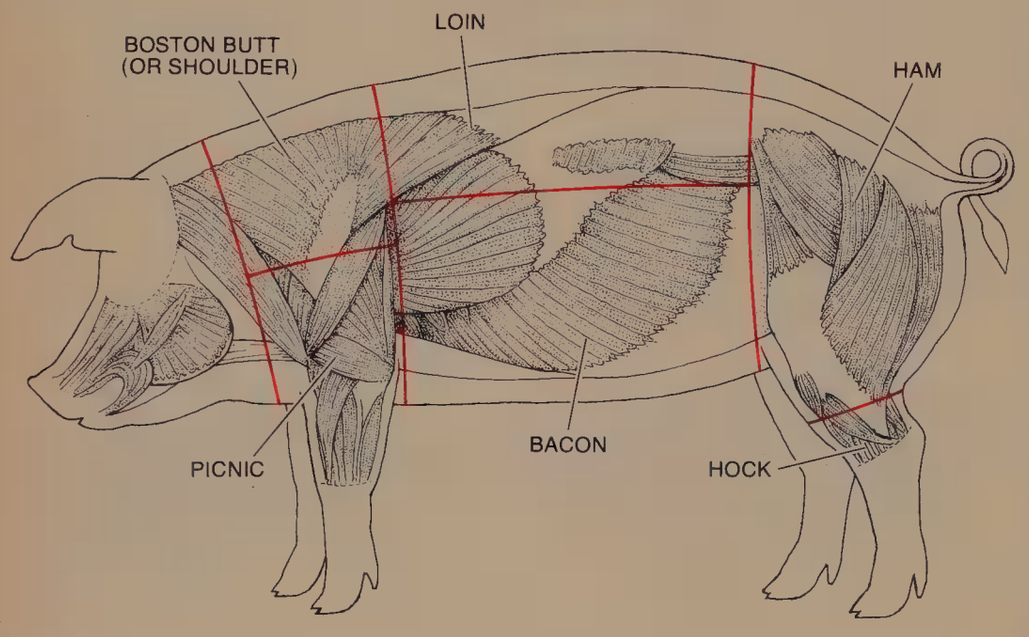
Crossbreeding, involving from two to four breeds in a typical systematic breeding program, accounts for more than 90 percent of the total annual supply of pork. Crossbred pigs are superior to their purebred parents in such economically important traits as rate of survival, rate of growth, efficiency in utilizing feed, production of milk and number of offspring. Crossbreeding is a valid method of production only if it continues to result in heterosis ("hybrid vigor"): the improved average performance of the crossbred progeny over that of their purebred parents. Continued progress in increasing the efficiency of swine production by crossbreeding depends on the intensive selection of purebred stock for excellence in the appropriate heritable traits.

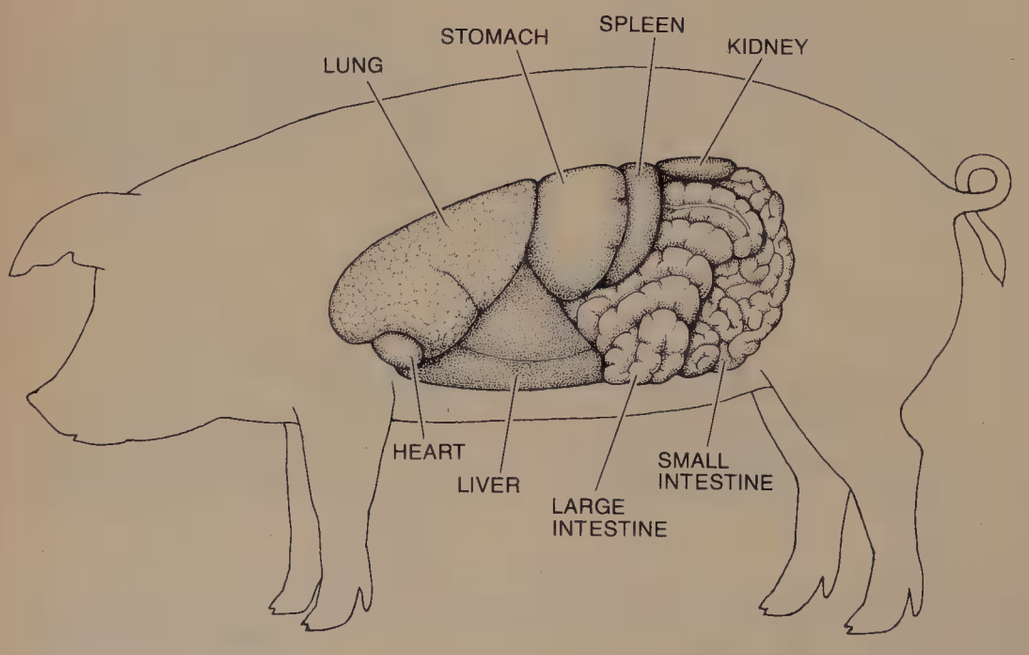
Some of the pig farms in the U.S. are "farrow to finish" enterprises, in which



EXTENT OF PIG PRODUCTION in various countries is suggested by this comparison of human and pig populations. For each country the black numeral is the human population, the colored numeral the pig population, both in millions. The bar shows the number of pigs as a percentage of the human population. The comparison is somewhat inexact because of variations in the "extraction rate," the ratio of the number of pigs sent to market per year to the number on hand at any one time. For example, in the U.S. the rate is about 1.5 and in Brazil about .4.







ANATOMY OF THE PIG is depicted in a presentation that shows the skeleton, the musculature and the internal organs. A widely employed butcher's terminology for cuts of the meat is also shown. Variations of and additions to the terms (chitterlings, for example) are often seen.

the pig spends its entire life on one farm. Others are "feeder pig" operations that raise piglets to the age of from eight to 10 weeks (shortly after weaning) and then sell them to "finishing pig" enterprises that feed them to marketable weight. The two types of operation that handle piglets do certain things routinely: clip the "needle teeth," or incisors, which might otherwise injure the sow when the piglet suckles; inject iron; disinfect the navel; notch ears for purposes of identification, and dock the tail because pigs have a tendency to chew the tails of other pigs.

Pig farms also maintain vaccination programs to control infectious diseases, spraying and drenching programs to control internal and external parasites, mating and culling programs to maintain reproductive efficiency and to optimize the use of the farm's facilities, and marketing programs to maximize the net income. A farm with purebred animals must in addition keep detailed and accurate records of the genetic identity of its pigs. It must also develop effective advertising and marketing programs so that other pig farmers will be aware of what the farm has to offer for crossbreeding. A finishing enterprise has fewer critical requirements in husbandry and is less labor-intensive, but it has to be aware of the condition of the market so that the farm does not have too many or too few pigs at a given time.

A major task in managing a pig farm is looking after the health of the animals. Pigs are vulnerable to a large number of infectious, metabolic and nutritional diseases. Modern advances in the prevention of infectious disease and in the control of external and internal parasites have reduced the veterinary bills the pig farmer once paid for the treatment of acute illness among his animals. Such advances have brought about increased emphasis on the health of the herd and on disease-prevention programs involving vaccination against infectious diseases, routine control of parasites and close restrictions on the movement of pigs from one farm to another. Many commercial pork-production enterprises maintain herds that are relatively free of disease by measures that include limiting human and animal traffic and quarantining new breeding stock brought into the herd; in the jargon of the industry such a herd is SPF, or specific-pathogen-free.

Certain metabolic diseases still cause significant financial losses even in well-managed herds. One is porcine stress syndrome, which is associated with specific breeds and genetic groups. Another is the metritis-mastitis-agalactia complex, a multifactor disease syndrome resulting in a high rate of death among newborn pigs.

Nutritional diseases are less common now than they once were, but a few still

appear even with today's sophisticated techniques for formulating diets. For example, an acute deficiency of selenium and vitamin E causes excessive deaths among growing pigs in certain regions of the U.S. and elsewhere that have a low content of selenium in the soil, resulting in a deficiency of the element in the plants grown there. As such deficiency syndromes are recognized the feed industry responds by providing appropriate supplements. Sometimes a supplement cannot be added to feed in the U.S. without the approval of the Food and Drug Administration. Selenium is an example, since it was known for its toxic properties before it was recognized as an essential nutrient; the FDA authorizes the addition of inorganic selenium to the diet of young pigs at levels of up to .3 part per million.

Prospective Changes

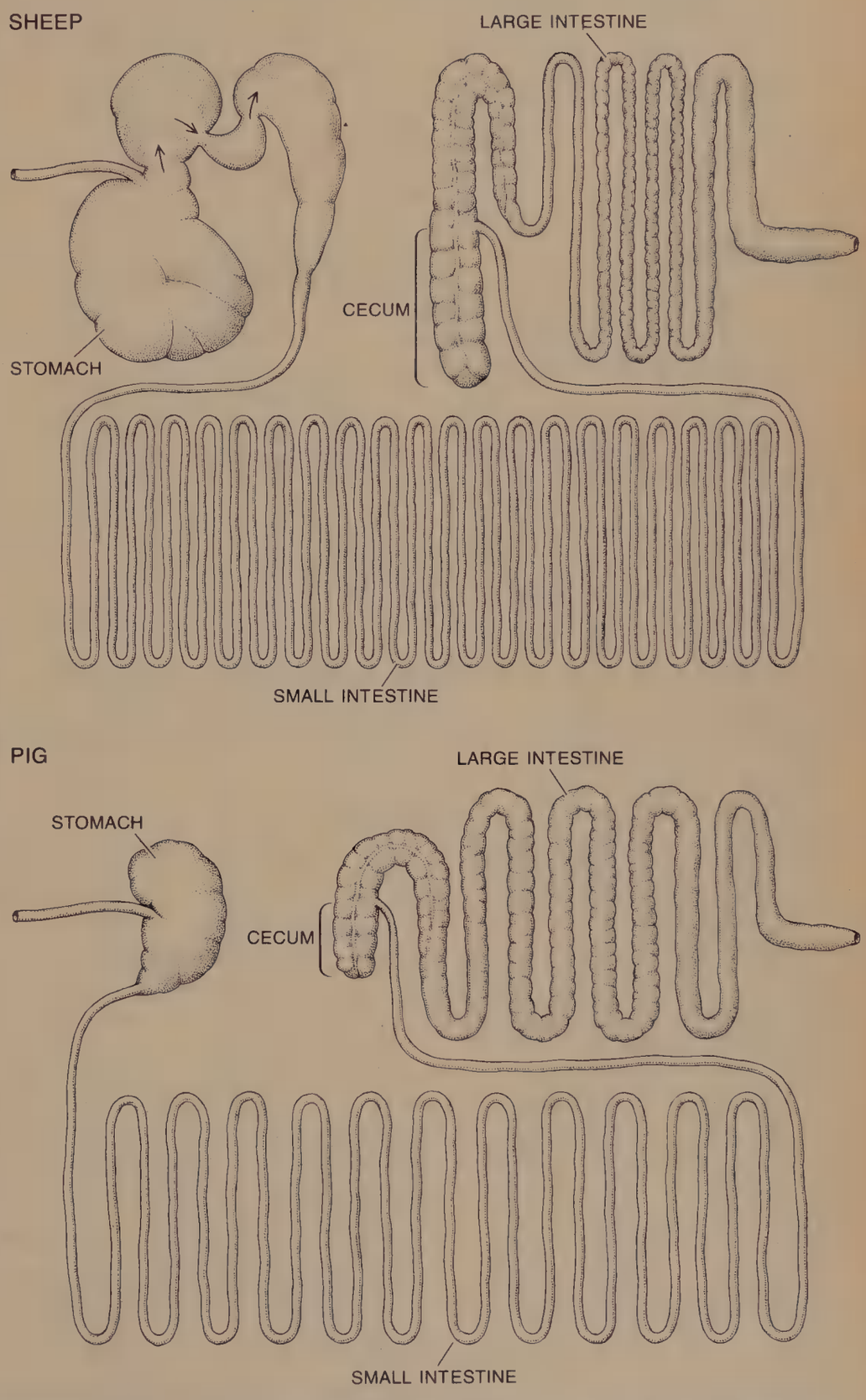
Although the pig industry has been greatly transformed over the past 30 years by advances in agricultural technology, biological knowledge and veterinary medicine, it seems probable that biological research now in progress will result in even further changes. They can be considered under the headings of genetics, nutrition and reproduction.

In genetics the farmer seeking better heritable traits in his crossbreeding activities is being aided by the work of population geneticists, who not only capitalize on the sophisticated modeling techniques made possible by computer technology but also introduce genetically diverse breeding stock into crossbreeding programs. Much research by geneticists in the U.S. Department of Agriculture and private organizations is directed at improving recognized traits that are beneficial in the production of pork and at finding new traits that might be beneficial.

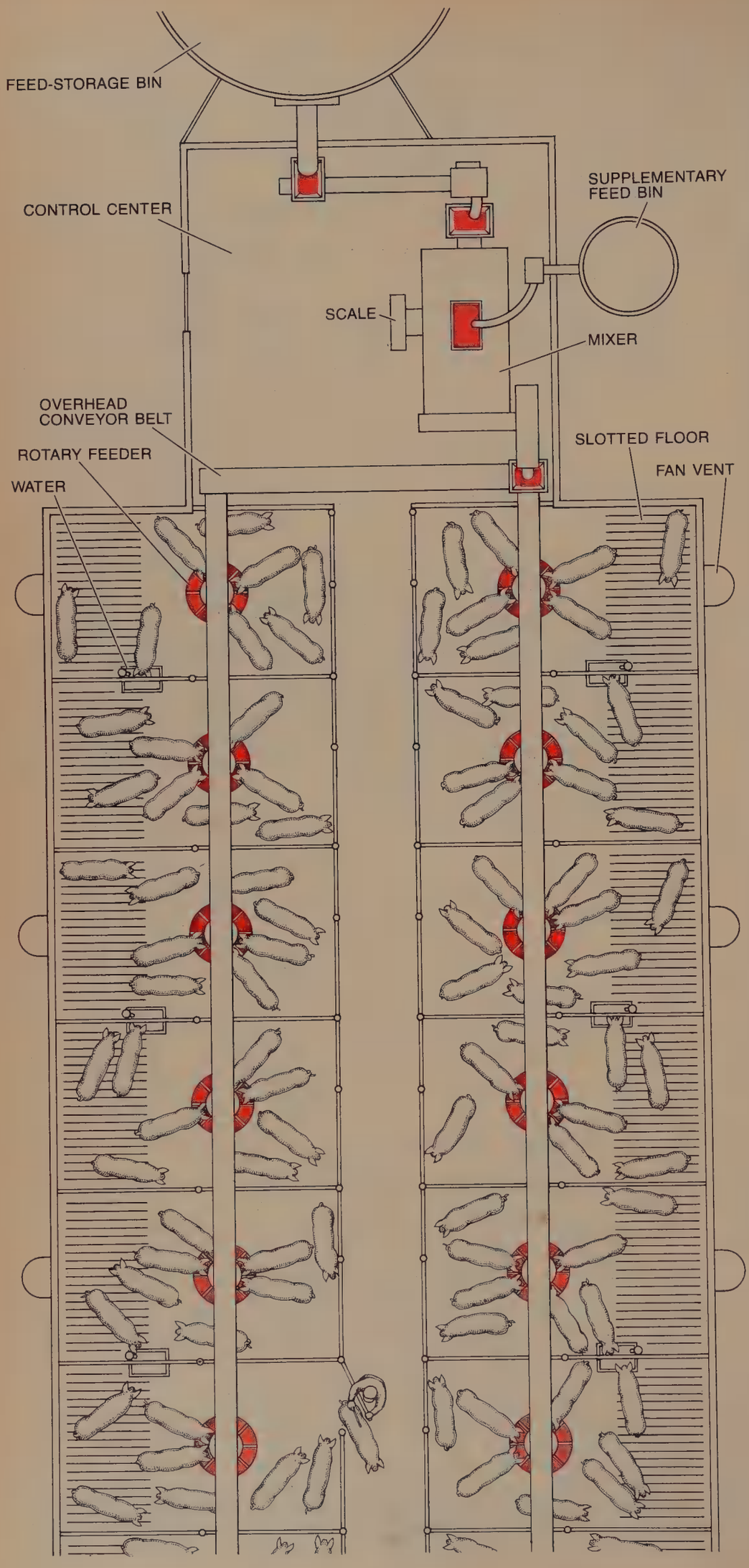
Techniques for estimating accurately the body composition of live animals, such as the measurement of back fat by ultrasonic devices, and for predicting the subsequent body composition of an animal by appropriate measurements done early in life offer possibilities for more effective selection of desired traits such as leanness. Genetic engineering by recombinant-DNA techniques ("gene splicing") shows promise of opening the way to the large-scale production of pig hormones by bacteria. With growth hormone made in this way the farmer might get his pigs to market weight sooner. Synthetic growth hormone is also being investigated as a modulator of milk secretion and the growth of lean tissue in a number of farm animals, including the pig.

Genetic variation in economically important traits such as growth and lactation is related to hormones, enzymes and various intracellular processes. The genetic bases for these relations are not well understood. Laboratory work on the growth of muscle cells in culture, together with other in vitro efforts to understand how protein and fat are synthesized and broken down, offers the possibility of fuller knowledge of the metabolic processes that control the buildup of muscle tissue in pigs and other animals.

In nutrition recombinant-DNA techniques may soon be applied to change the activity of the microflora in the pig's gastrointestinal tract in order to make



DIGESTIVE SYSTEMS of the sheep and the pig are compared. The sheep is a ruminant, with a large stomach consisting of four parts: the rumen, the reticulum, the omasum and the abomasum. The system evolved to digest cellulose. Grass eaten by the sheep passes first into the rumen; it is later regurgitated and chewed further. After it is swallowed again it passes successively into the other three parts of the stomach. The pig is an omnivore, but it cannot digest much cellulose. Research under way aims to exploit recombinant-DNA techniques ("gene splicing") to change the activity of the microflora in the pig's gastrointestinal tract to deal with preparations derived from grass and other fibrous plants, greatly expanding the sources of pig feed.



cellulase, the enzyme that cleaves cellulose, the most abundant constituent of plants, into its glucose monomers. Such a development would change radically the constraints on pig production imposed by the inability of the animal to digest cellulose and other complex carbohydrates. Foods the pig cannot now eat or fully digest would then be available to pig husbandry.

Recombinant-DNA techniques applied in another important way may soon make amino acids, the monomers of protein, available to animal husbandry. Recombinant-DNA strains of bacteria that are efficient producers of amino acids can be grown on a pilot scale. If amino acids made in this way were available at prices competitive with the cost of standard feed, more than four billion pounds of feed protein could be saved annually. The metabolic needs of the farm animal could be met with a low-protein feed combined with a well-balanced mixture of amino acids.

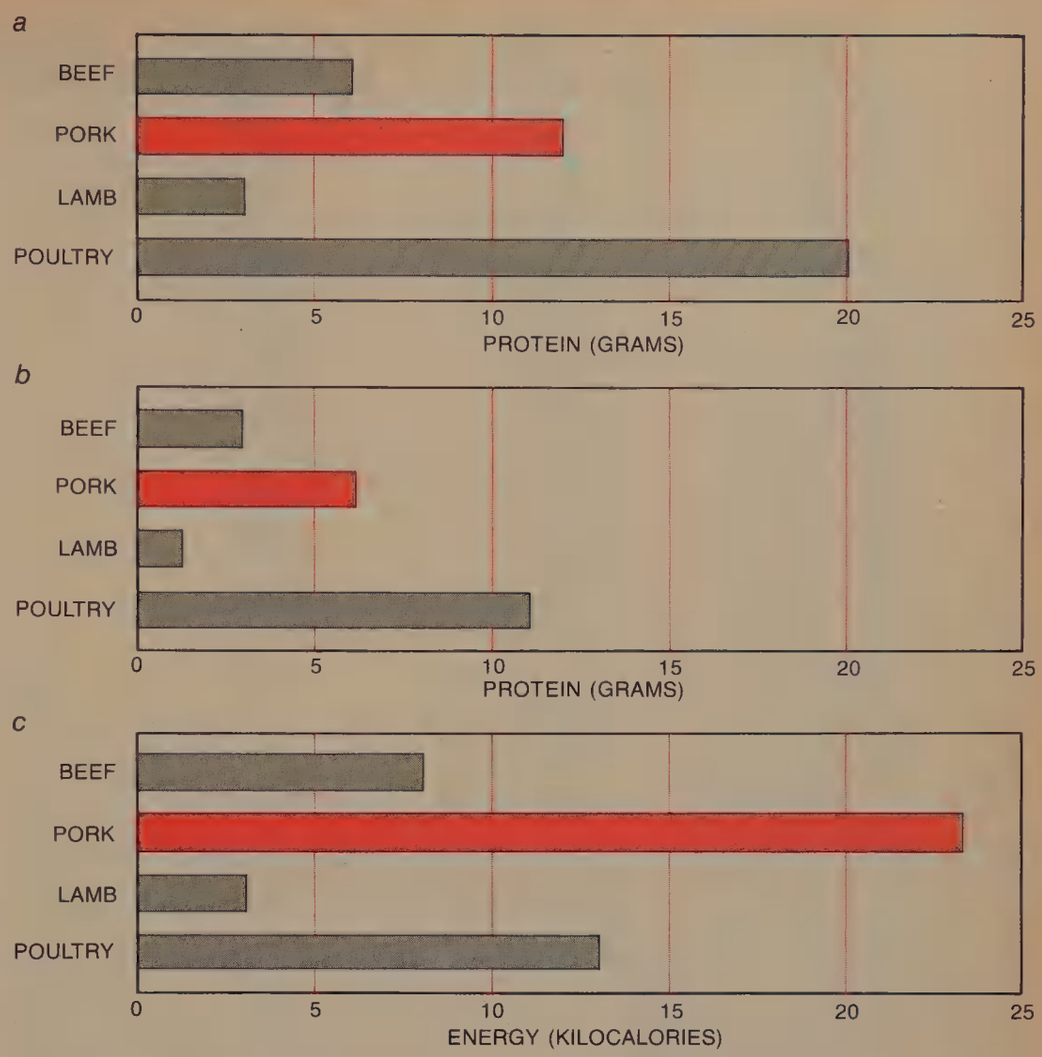
New types of feed will also play an important role in the future of pork production. Plant breeders continue to develop new varieties of corn, barley and oats with a high content of lysine, an amino acid essential to animal nutrition that is deficient in older varieties of these grains. Animals fed with the high-lysine varieties would need less supplemental protein. In developing countries pigs are being fed local products that were not traditionally a part of the pig's diet; examples include sweet potatoes and fish meal in New Guinea and bananas in Ecuador. Finally, several microbial sources of protein, including dried bacteria, yeast and algae, offer a potential supply of feed for pig farmers.

The interest in improving reproduction stems from the high death rates among pigs before birth and soon afterward. In the U.S. prenatal mortality is normally from 20 to 40 percent. A sow releases about 17 ova per period of estrus. Although about 95 percent of the eggs are fertilized, only 9.4 pigs are born on the average and 7.2 survive to marketable age. Even in Denmark, where emphasis is laid on efficiency in pork production, 20 percent of the pigs in a litter die before being weaned at the age of from six to eight weeks.

AUTOMATIC FEEDING ARRAY for pigs is designed to be part of a total-confinement unit. The feed from the storage bin is corn or cereal grains, moved from the bin to the mixer by an electrically operated auger. Protein, mineral and vitamin supplements are added at a rate of about one pound for each four to eight pounds of grain. The mixture is moved to the feeders by overhead conveyor belts or augers. The slotted floor facilitates waste removal. The feeder portrayed on the cover of this issue is a unit in such a feeding system.

Research on the genetic, endocrine and nutritional factors associated with high prenatal mortality shows the promise of making it feasible to induce superovulation—the release of a larger number of eggs—by administering suitable hormones. It has been possible for some time to synchronize ovulation in a group of virgin females by administering hormones, but the practice has not been widely followed because techniques for freezing the semen of boars have not been available. Synchronized ovulation or superovulation coupled with artificial insemination with fresh or frozen semen from genetically superior boars may make it possible to achieve higher rates of live births per pregnancy.

An improvement in reproductive efficiency must be accompanied by improved methods of caring for piglets in the days following birth. The sow's colostrum, the milk she secretes in the first few days after she gives birth, contains antibodies against pathogens that may infect the piglet. It has been found that after the piglets have received the colostrum for two or three days they can be reared successfully on liquid or dry diets. Indeed, dry diets may be associated with higher survival rates and a lower incidence of diarrhea. Combining the technology that would raise the average of viable pigs per litter to 15 or 20 with the early weaning of part of the litter after two or three days could result in a significant increase in the number of pigs marketed annually per breeding female from the present norm of from 12 to 14 to an average of 24 or more.



CONVERSION EFFICIENCY OF FOUR MEAT ANIMALS is charted. Shown are the amount of edible protein for each 100 grams of food consumed by the animal (*a*), the amount of edible protein per 100 kilocalories of metabolizable energy the animal eats (*b*) and the kilocalories of edible energy per 100 kilocalories of metabolizable energy in the animal's food (*c*).

Attainable Goals

The prospective advances of pig farming imply it is possible to envision certain goals that should be attainable within the next 20 to 30 years. This has been done by workers at universities and by my colleagues at the Meat Animal Research Center of the Department of Agriculture for eight factors that are economically important in the production of pigs. The number of litters borne per sow annually should rise from 1.8 to 2.2 and the number of pigs marketed per sow annually from 13 to 24. Today's high losses should decline; expressed one way, as the percentage of loss of pigs between birth and weaning, the change should be from the current 15 to 30 percent to 5 to 10 percent, and expressed in terms of the number of pigs raised per litter the change should be from 7.2 to 11. The number of pounds of feed per pound of gain in body weight from birth to market is now 3.5 and can be expected to be 2.5 or less, and the months of age from birth to market should decline from the five or six of the present to four or five. Because leanness of meat is increasingly sought by consumers, one can expect the percentage of fat in the edible parts of the carcass to decline from 41 percent to 32 percent or less and the average thickness of back fat at the time of marketing to decline from 1.6 inches to less than .8 inch.

In 1978 Henry A. Fitzhugh and some of his colleagues at the Winrock International Livestock Research and Training Center projected the proportions of food energy and protein that human beings will obtain from pigs, ruminants and poultry by the year 2000. They foresaw an increase of 23 percent in the number of pigs, 23 percent for poultry (chickens, ducks and turkeys) and 29 percent for ruminants (cattle, sheep and goats) between 1970 and 2000. In addition they predicted that food energy from pigs will be 215 billion megacalories in 2000 compared with 144 billion in 1970 and that protein from pigs will be three million metric tons compared with two million in 1970. Pork will provide 23 percent of the energy and 8 percent of the protein from animal sources, including milk, compared with 25 and 9 percent respectively in 1970. In other words, although the production of pork is predicted to increase substantially, the overall contribution of pork to human nutrition is predicted to decline slightly.

Other Roles of the Pig

Besides its importance to agriculture the pig is valuable as a model animal for many physiological and nutritional studies related to human health. Pigs are similar to human beings in the anatomy and physiology of several organs and organ systems, including the teeth, the eyes, the kidneys, the cardiovascular system, the skin and the digestive system. Studies of the pig have helped in elucidating a variety of human disorders such as cardiovascular disease, diabetes, gastric ulcer, muscular dystrophy, alcoholism and obesity.

In clinical medicine the heart valves of the pig have often been inserted as a replacement for damaged human heart valves. Several vital substances extracted from the tissues of the pig are widely employed in medicine. They include insulin from the pig pancreas and heparin (an anticoagulant) from the pig lung. Moreover, the skin of the pig can be grafted successfully onto the human body to repair burns.

New Inorganic Materials

Synthetic organic materials have come to play a key role in modern civilization, but unusual materials made out of sand, clay and other minerals have the advantage of requiring a smaller input of energy

by J. D. Birchall and Anthony Kelly

Look around you and you are likely to see things made of a wealth of diverse materials: metals, plastics, ceramics, glass and concrete. They are the stuff of a technological society. It takes energy to make them. In fact, the worldwide production of materials for construction and manufacturing annually calls for the equivalent in energy of 10^9 tons of oil, or 15 percent of the total worldwide expenditure of energy. The amount is equal to the total worldwide expenditure of natural gas.

The amount of energy required to make a material varies greatly from one material to another. It is useful to take portland cement as a standard, because it is made from chalk and clay, two readily available constituents of the crust of the earth. Moreover, portland cement is produced on a huge scale: almost 10^9 tons per year worldwide. It takes about 3×10^{10} joules of energy to make a cubic meter of portland cement. It takes six times as much to make a cubic meter of polystyrene plastic and 29 times as much to make a cubic meter of stainless steel.

Plainly portland cement, an inorganic material, has a considerable advantage in energy over plastics, which are made from organic material, and metals, which are smelted by processes that require very high temperature. This advantage, together with the wide and ready availability of the earth chemicals (the oxides of silicon and aluminum, the carbonates of calcium and magnesium and so on), suggests that technology should reexamine inorganic materials as replacements for energy-expensive materials, particularly at a time when energy itself is expensive and an end can be seen to the ready availability of hydrocarbons in the form of oil. Research we shall describe in this article is demonstrating that inorganic materials can in many applications indeed supersede the energy-expensive materials of present-day technology.

The first materials employed by human beings were flint and other kinds of stone, along with materials of animal or vegetable origin such as skin, bone and wood. Together they enabled human beings to excel at hunting and food gathering. Flint is hard, brittle and amorphous in its microscopic structure; thus it fractures conchoidally and can be fashioned into a sharp cutting edge. The Stone Age use of flint grew to such a scale that flint had to be mined rather than searched for in surface deposits.

The next material was bronze, which must have arisen from the chance juxtaposition of easily smeltable rocks (ores of copper and zinc) with sources of heat and with carbon. At much the same time it was discovered that wet clay would harden when it was baked and that before the baking it could be molded into a useful shape. Somewhat later came the reduction of iron ore into metallic iron. The earliest synthetic-material technologies, then, were metallurgy and ceramics.

A much later advance was the discovery, by the Romans, that a mixture of volcanic ash and lime reacts with water to form a hard, solid mass: a cement. This was the first low-temperature inorganic technology. In the reaction the silica (SiO_2) in the ash combines with the lime ($Ca(OH)_2$) to form a calcium silicate that has no regular chemical composition. In 1824 Joseph Aspdin improved on the technology by heating lime with a material such as clay, which contains silica. The result, after cooling, is portland cement clinker, which includes the silicates $2CaO \cdot SiO_2$ and $3CaO \cdot SiO_2$. When the clinker is mixed with water, it sets to a hard, firm, rock-like solid. Instead of shaping rock by chipping or cutting, civilization could now cast or mold it into the familiar form of concrete (cement mixed with sand and stone).

Metals, ceramics, rock and cement continued to dominate technology until the early years of this century. Cement, however, served only in low-grade applications. Then came the age of plastics, which can be said to have begun in 1907, when Leo H. Baekeland produced the first synthetic resin, Bakelite. Even before that, some natural polymers had been modified into materials with novel and useful properties. Natural rubber, for example, stretches when it is pulled. In 1839 Charles Goodyear found that the addition of sulfur stiffens it. The sulfur atoms form bridges between the hydrocarbon chains of which the rubber is composed. Nitrocellulose, developed by Christian Friedrich Schonbein in 1845, is brittle. In 1864 Alexander Parke discovered that the addition of camphor, a gummy substance from the camphor tree, makes it pliable. The camphor molecules act as a lubricant between hydrocarbon chains. The plasticized nitrocellulose served as the first motion-picture film.

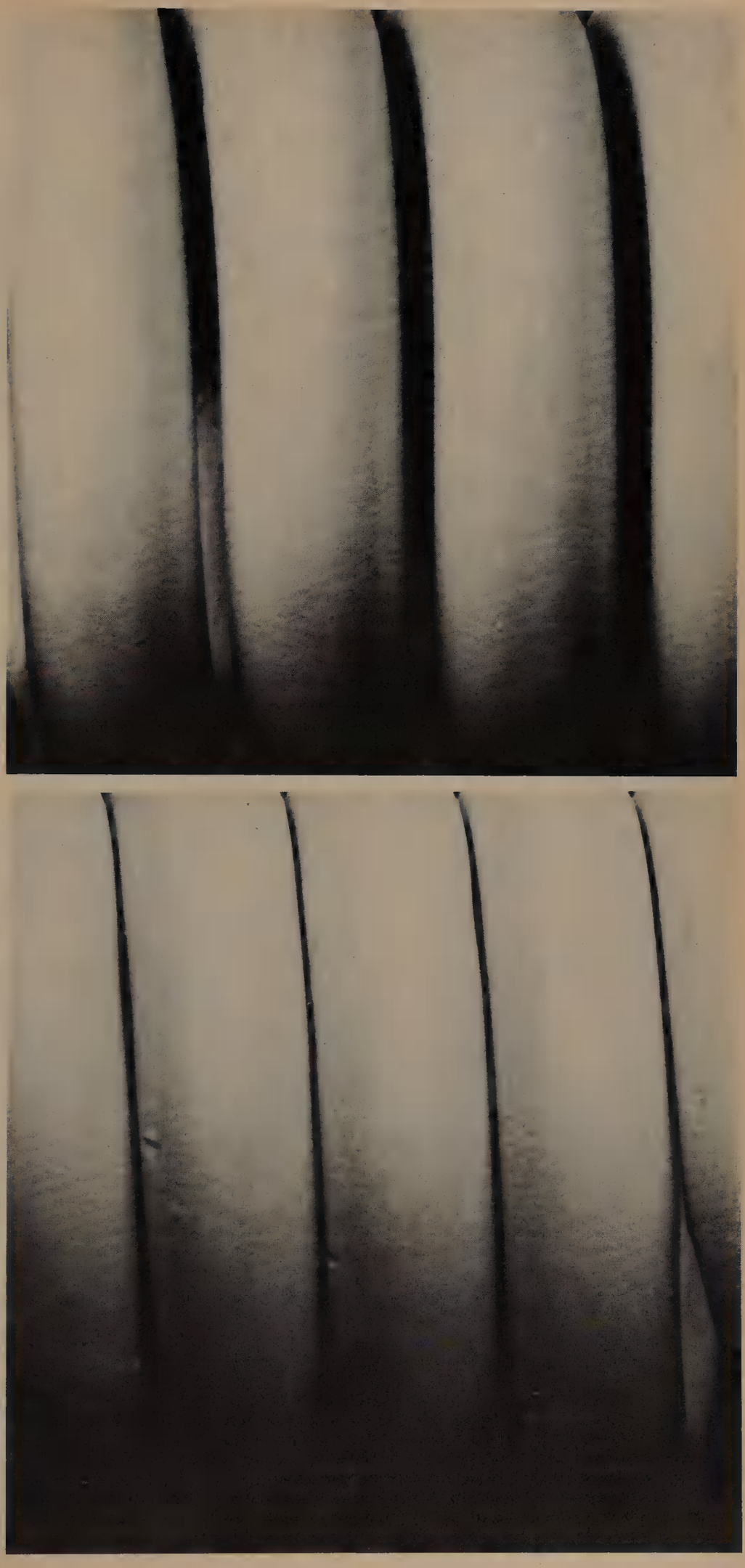
In the wake of Bakelite the age of plastics began in earnest. Synthetic rubber appeared in 1910, polystyrene in 1925 and nylon, polyethylene and polyesters in the 1930's. By 1940 the synthetic-polymer industry was well established. In that year the raw material for some 95 percent of all synthetic organic chemicals was coal. Today, four decades later, when the production of synthetic organic chemicals has increased more than a hundredfold, the raw material for some 97 percent is petroleum. Although the price of petroleum is currently decreasing somewhat, over the past eight years it has increased some fifteenfold.

What is the future for materials technology? One prospect is to derive synthetic polymers from coal once again. Surely that will happen. Another prospect is that the earth chemicals will again become prominent as raw materials. Apart from their availability, inorganic solids are intrinsically much stiffer than organic polymers. Moreover, concern is growing about the fire hazard of organic materials in homes and public buildings. Inorganic materials do not burn.

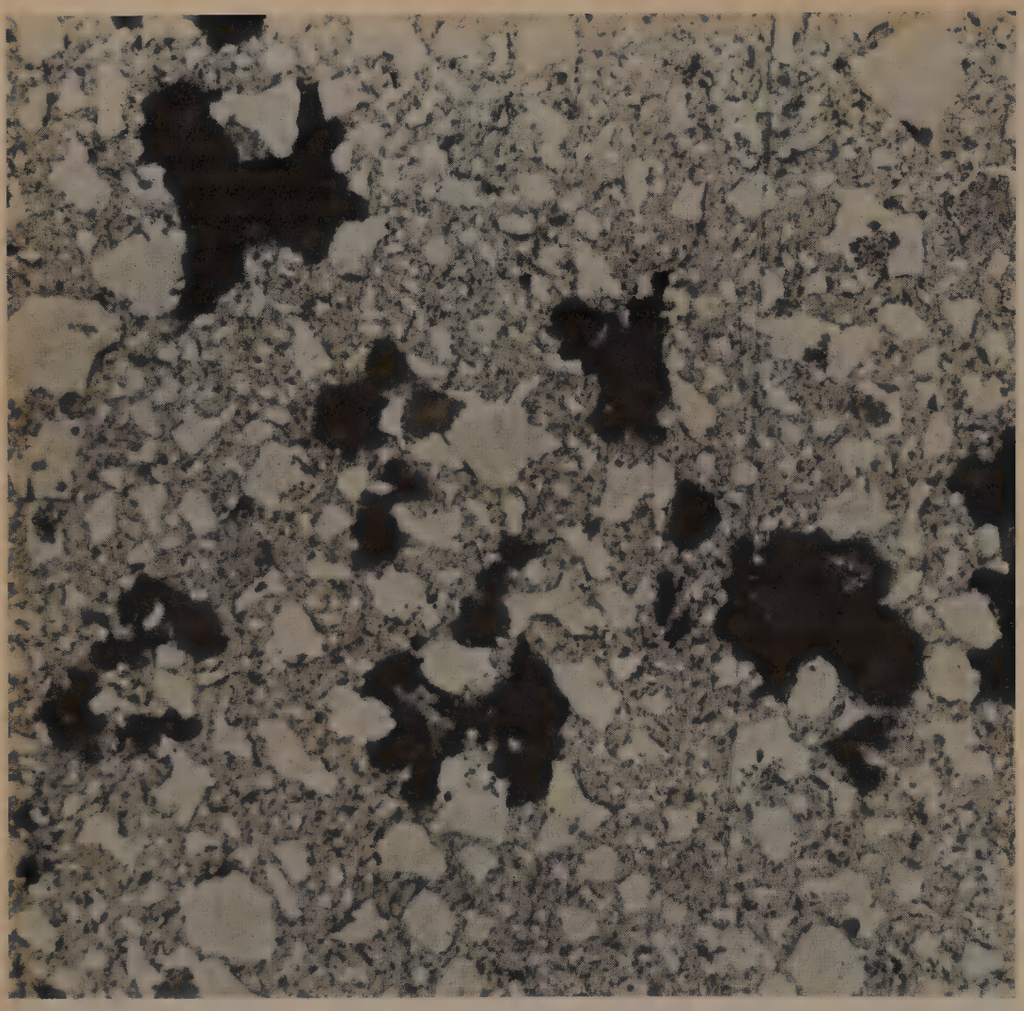
In the 1960's it was hoped that inorganic polymers would be synthesized in great variety and would find commercial application. Each inorganic polymer would be analogous to the organic polymers in being a long chain of repeating monomers, but the backbone of the chain would consist of atoms other than carbon. To date the only successful synthetic inorganic polymer that has emerged is silicone. In silicone the backbone consists of silicon atoms and oxygen atoms in alternation. Hence synthetic inorganic polymers have largely remained a dream. On the other hand, inorganic polymers taken fairly directly from nature are very widely used indeed; an example is glass. Moreover, the silicate rocks of the earth's crust consist of both chains and rings of silicon atoms and oxygen atoms in alternation. The problem is that the production of materials such as glass, as it is currently conceived, calls for high temperature, and high temperature defeats the objective of turning to inorganic materials in order to conserve energy.

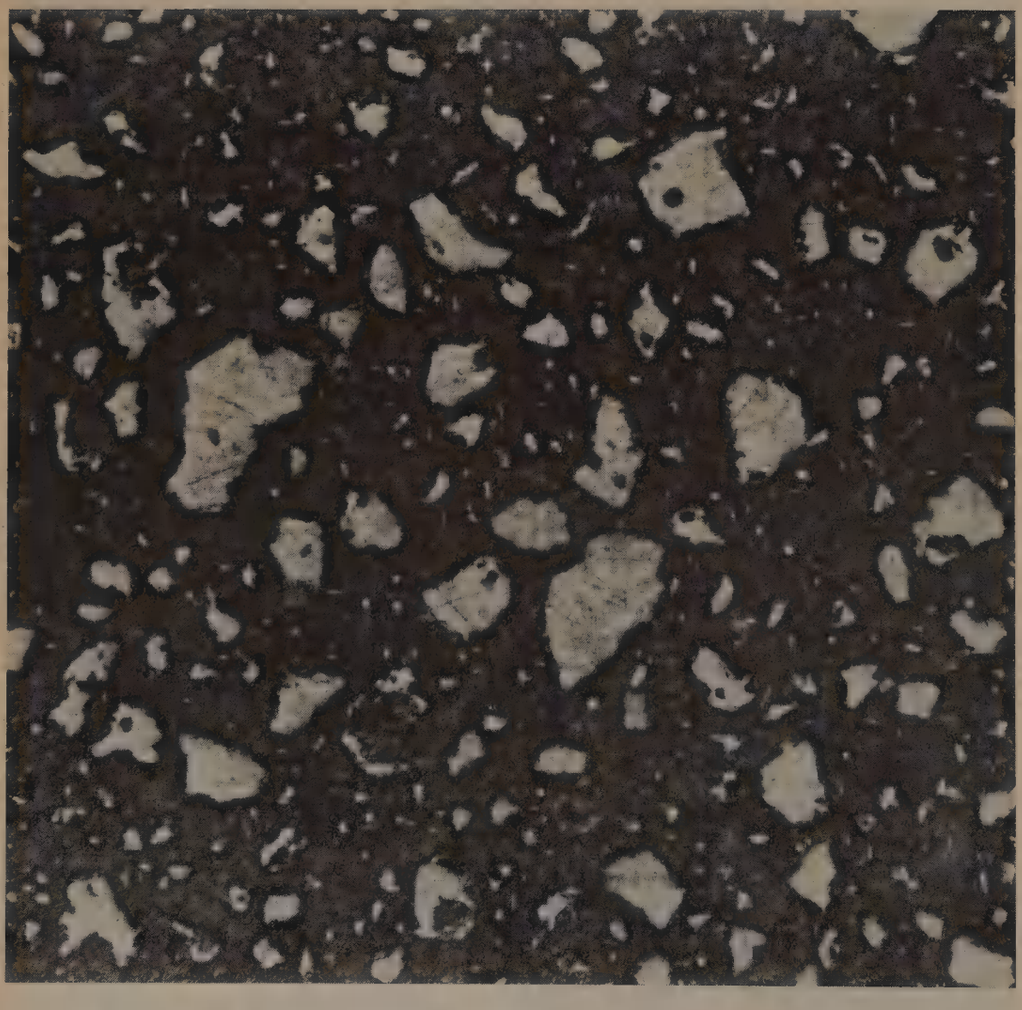
There is a second problem. Consider the basic properties of materials that determine how the engineer incorporates them into a design. Three such properties are stiffness (the resistance of a material to bending), tensile strength (resistance to pulling) and fracture toughness (resistance to impact). They are properties any do-it-yourselfer would judge informally as he handled a material and decided whether or not to use it. Organic polymers and organic materials such as wood are not as stiff as inorganic materials such as ceramics and glass (one reason many synthetic organic materials are reinforced by incorporating into them fibers of inorganic material). The inorganic materials, however, lack toughness. If inorganic materials are to stand any chance of replacing plastics (and metals) in some applications, their toughness will have to be improved.

Can these two problems be solved? That is, can sufficiently tough inorganic materials be made at low temperature? The study of living organisms suggests they can. Living organisms exploit calcium carbonate, calcium phosphate and silica with remarkable control and sophistication. The abalone shell, for example, is 99 percent calcium carbonate; it is almost entirely chalk. Yet its tensile strength is more than 100 megapascals (14,500 pounds per square inch) and its



SPRING MADE OF CEMENT is shown under 300 pounds of tension (*top*) and under no tension (*bottom*). The cement is specially prepared to have no pores or other internal flaws larger than a few micrometers across. As a result the cement, which is called macro-defect-free, or MDF, cement, is highly resistant to fracture. Indeed, it is as strong as aluminum.





ABSENCE OF LARGE PORES in MDF cement is displayed in micrographs of a polished section of ordinary cement (*top*) and MDF cement (*bottom*). The pores in ordinary cement (*black patches*) can compose as much as 30 percent of the volume of the material. They are responsible for its lack of strength. The micrographs were made at an enlargement of 50 diameters.

toughness is more than 1,000 joules per square meter of new surface generated by fracturing, which makes it comparable in that respect to polystyrene and Plexiglas.

An examination of the shell's microscopic structure by means of electron microscopy shows how these properties are achieved. Flat plates of calcium carbonate, each about .2 micrometer thick, are stacked with impressive uniformity and are separated from each other by a thin layer of protein. A crack in the shell is thus forced to traverse a tortuous path and so the shell's fracture toughness is great. In a way that is not yet understood the protein serves the abalone as a template for the deposition of an ordered pattern of inorganic material at the temperature of the environment. The same is true of the growth of bone.

Let us examine in detail how a man-made inorganic material, say the ceramic porcelain, is created. The contrast with how the abalone fashions its shell is instructive. The basic process of making porcelain was discovered in China and was brought to the West in the early 18th century by Father F. X. d'Entrecolles, in what was probably the first recorded instance of industrial espionage. d'Entrecolles had spent time in Jingdezhen, a center of porcelain making in Jiangxi Province, and had learned Chinese in order to determine the role in porcelain making of the kind of clay the Chinese called kaolin. (To this day kaolin is one of the few Chinese words in the English language.)

High-quality ceramics such as porcelain are made by firing, that is, by heating to high temperature a mixture of quartz, feldspar and clays (usually kaolin and ball clay, a fine-grained clay that has some organic matter in it). To begin with, each of the powdered raw materials is mixed with water. The resulting pastes are pressed in a cloth to remove the excess water; then the pastes are combined and extruded through a screw under vacuum to remove most of the air in the mixture. The mixture is dried and shaped, much as a potter would shape it, and the resulting object is put into a gypsum mold to remove still more of the water in it. Next the object is coated with a glaze. The glaze is usually a clay combined with an additional flux: a material that readily fuses.

After further drying the object is fired. At temperatures of some 400 degrees Celsius a number of chemical reactions begin. Then at 600 degrees the material becomes glassy and crystals form in it, so that the material shrinks. The temperature is only about a third of the material's melting point, but that is deliberate: melting the material would give rise to large, irregular crystals rather than tiny, regular ones and in any case would re-

NOW, THE JOY OF GREAT SETS

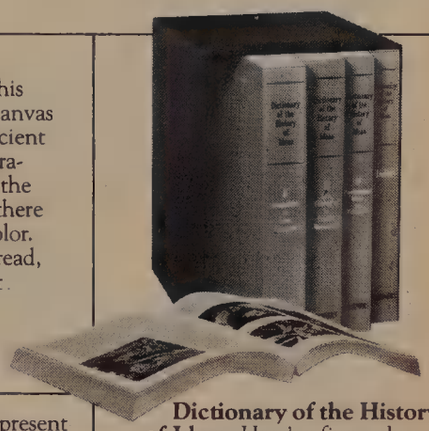
Great sets make for good reading. Now, the biggest sets offer in QPB history also makes for great savings. Enjoy up to $262.05 off the hardcover price when you choose one of these sets.

QPB books are soft-covers in hardcover sizes, durably bound and printed on fine paper. But they cost up to 65% less than their hardcover counterparts.

And for the first time QPB is giving you a chance to get your books on tape, too.

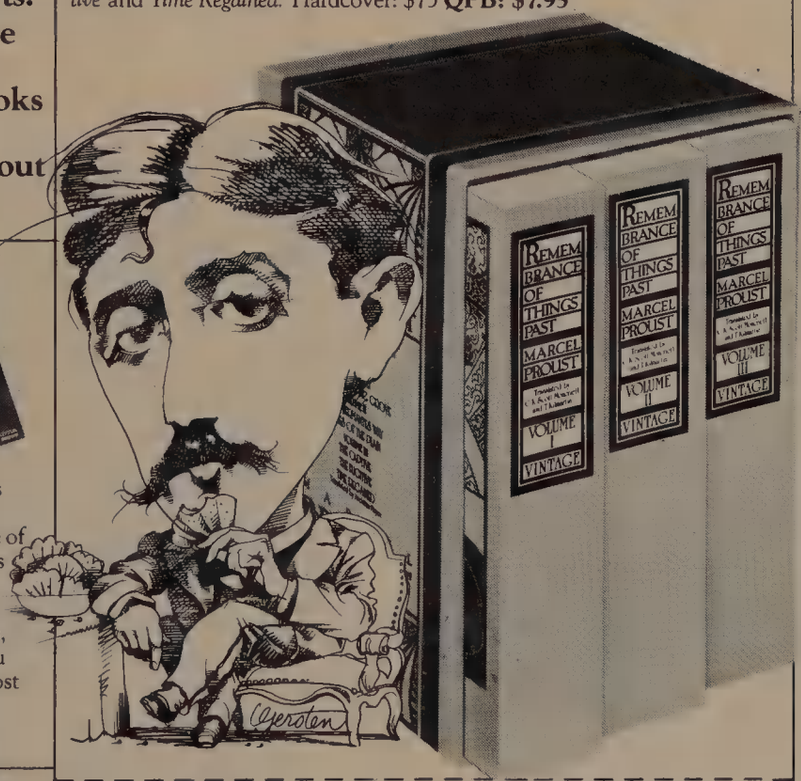
Join QPB, and find out how great sets can be.

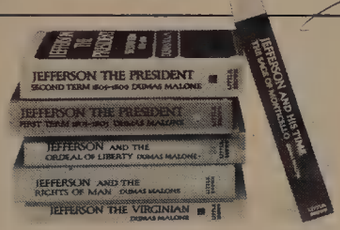
World of Art Library. This nine-volume set lets you canvas the history of art, from ancient Greece to modern photography. Elegantly printed for the Oxford University Press, there are 1848 plates—360 in color. Authoritative yet easy to read, this set is a picture-perfect choice. Publisher's prices total: $86.55 **QPB: $12.50**



Dictionary of the History of Ideas. Here's a five-volume illustrated set filled with ideas. 311 articles by 254 respected authorities give you the essential ideas of Western thought. Cross-references and bibliography follow each article, so you can follow ideas as they develop through the centuries. Can you think of a wiser investment? Hardcover: $275 **QPB: $12.95**

Remembrance of Things Past. There's no better time than the present to read Proust's *Remembrance of Things Past.* That's why this acclaimed translation is creating a new audience for Proust's masterpiece. Handsomely boxed, this three-volume set (3365 pages, with notes and synopsis) includes all seven parts of this modern classic—*Swann's Way, Within a Budding Grove, The Guermantes Way, Cities of the Plain, The Captive, The Fugitive* and *Time Regained.* Hardcover: $75 **QPB: $7.95**





Jefferson and His Time. Dumas Malone's Pulitzer Prize-winning biography has been hailed as "one of the most monumental biographies of our time" by *The Washington Post.* Six volumes, encompassing 3346 pages and many illustrations, this definitive biography helps you find out more about America's most remarkable founding father. Hardcover: $121.70 **QPB: $8.95**

Charles Dickens: Readings. Reading from Dickens out loud was once an American pastime. Today, you can listen to the best of Dickens on four QPB tape cassettes. James Mason reads selections from *A Tale of Two Cities,* Roger Rees (of Broadway fame) reads from *Nicholas Nickleby* and Sir Lewis Casson and Boris Karloff read from *The Pickwick Papers.* Get the Dickens cassettes, and hear what classics are all about. Manufacturer's suggested list prices total: $35.92 **QPB: $7.50**

Let's try each other for 6 months.

Quality Paperback Book Club®, Middletown, Pa. 17057. Please enroll me in QPB and send me the set I have checked below, billing me for the appropriate amount, plus shipping and handling charges. I understand that I am not required to buy another book. You will send me the *QPB Review* (if my account is in good standing) for six months. If I have not bought and paid for at least one book in every six-month period, you may cancel my membership. A shipping and handling charge is added to each shipment.

Check one box only
☐ 900. Dict. of Ideas $12.95
☐ 901. World of Art $12.50
☐ 902. Dickens $7.50
☐ 903. Remembrance of Things Past $7.95
☐ 904. Jefferson $8.95

Name _____ 3-11

(Please print clearly) QB 170-5

Address _____ **Apt.** _____

City _____ **State** _____ **Zip** _____

QPB is a Division of Book-of-the-Month Club, Inc.
Prices generally higher in Canada.

Join now. Save big on one of our great sets with no obligation to buy another book.

How membership works.

1. You receive the **QPB Review** 15 times each year (about every 3½ weeks). Each issue reviews a new **Selection,** plus scores of Alternates.
2. If you want the **Selection** do nothing. It will be shipped to you automatically. If you want one or more Alternate books—or no book at all—indicate your decision on the Reply Form always enclosed and return it by the date specified.
3. **Bonus books for Bonus Points.** For each softcover book or set you take (except for the books in this offer), you earn Bonus Points which entitle you to choose any of our softcover books. You pay only shipping and handling charges.
4. **Return privilege.** If the *QPB Review* is delayed and you receive the **Selection** without having had 10 days to notify us, you may return it for credit at our expense.
5. **Cancellations.** You may cancel membership at any time by notifying QPB. We may cancel your membership if you elect not to buy and pay for at least one book in every six-month period.

QPB®

The first book club for smart people who aren't rich.

quire a large and costly expenditure of energy. On the other hand, atoms diffuse slowly in an unmelted material, so that when water evaporates, pores are left behind. Moreover, holes remain unfilled where clay particles fail to pack tightly. Further still, the individual crystals in the material are of varying composition and so are quite different in their mechanical properties. Notably they differ in their coefficient of thermal expansion: the degree to which heating makes them expand and cooling makes them contract. When the ceramic is cooled, therefore, some of the pores and holes may fill up but cracks and other blemishes develop. When the ceramic returns to room temperature, it is brittle and weak.

That is why all useful ceramic articles are coated with a glaze. The glaze is chosen to have a coefficient of expansion lower than that of the ceramic itself. Hence the cooling that follows the firing of the ceramic puts the glaze into compression. In effect the glaze holds the ceramic in a kind of vise, making it difficult for cracks to enter the ceramic from its surface and thereby lead to breakage.

Paradoxically, then, the high-temperature process of firing, which is intended to give the ceramic strength by allowing diffusion and shrinkage to fill up the gaps in the material, succeeds only at the expense of introducing cracks. High temperature, however, is not the only way to make atoms move so that they fill up gaps. Another way is to make water serve as a solvent for ions and as a medium for their diffusion. That is what happens in the hydraulic cements: solids that set and harden irreversibly in the presence of water. Gypsum is one example; portland cement is another.

When the calcium silicates in portland cement are mixed with water, some of the water takes part in the production of hydrated calcium silicates. Even today, 150 years after the invention of portland cement, the reactions are not well understood. The rest of the water is lost by evaporation, leaving pores in the cement. The pores vary in diameter from a few tens of angstrom units to more than a millimeter; together they can occupy 25 to 30 percent of the volume of the solid.

It has been known for many years that the strength of a cement, like the strength of a fired ceramic, depends on this porosity. The exact nature of the dependence has been elusive. First it was found that the strength of a cement depends on the ratio of water to solid when the cement is made. The strength increases as the proportion of water is reduced. Then it was proposed that the strength depends on the volume of the

CEMENT BOTTLE CAP is an instance in which MDF cement replaces plastic. The cap has been fractured, and the fracture is shown crossing the thread of the cap in a scanning electron micrograph made at an enlargement of 75 diameters. The fracture is the dark area.

down to size.

(to an average of $98/month)

You can now reduce your student loan payments dramatically—by up to 50%—in the OPTIONS program at Sallie Mae.

Sallie Mae is a private corporation chartered by act of Congress in 1972 to make it easier for lenders to provide student loans. Now, we've been authorized to make student loans easier to pay off and we're doing just that with OPTIONS.

We simply pay off your existing NDSL, GSL, and FISL loans and create one new, 7%-interest, guaranteed student loan with no prepayment penalty. By choosing a longer payback period, payments that start low and rise gradually, or a combination of both, you end up with a single, lower monthly payment. The average reduction is from $151 to just $98 a month.

Loan Category	Typical Old Payment	Typical OPTIONS Payment
$5,001-$7,500	$ 96.00	$ 58.00
$7,501-$11,000	122.00	72.00
$11,001-$15,999	166.00	96.00
$16,000-or more	214.00	123.00

What OPTIONS can do for you. To get an idea of how much we can reduce your monthly payments today, estimate the amount you still owe on any NDSL, GSL, or FISL loans in good standing. Find the total on this chart and you'll see what we've done for borrowers like you.

With OPTIONS, you get to choose the payment plan that fits into *your* plans. When you get a student loan, most lenders tell you how much you'll pay, and for how long. With OPTIONS, *you* tell *us*. With OPTION 1, you can select a longer payback period, for a lower, unchanging monthly payment. With OPTION 2, payments start very low and increase gradually—like your income. OPTION 3 gives you graduated payments, too, but lets you pay off the loan faster.

If you have student loans to pay off, exercise your OPTIONS. OPTIONS is your opportunity to create the student loan payment plan that's right for you...to cut your student loan payments down to a size you can live with right now, when your income is probably lower than it will ever be again.

So if you owe more than $5,000 in NDSL, GSL, or FISL loans, don't miss this unique opportunity.

Let us give you more details, at no obligation. Thousands of student loan borrowers have already taken advantage of the OPTIONS program. To find out why, fill in and return the coupon. We'll send you a brochure that fully explains the OPTIONS open to you. Or, if you have any questions, call our toll-free 800 number.

Name_____

Address_____

City_____ State_____ Zip_____

SallieMae
OPTIONS

Mail to: B107
Student Loan Marketing Association,
OPTIONS, 1050 Thomas Jefferson St., N.W.
Washington, D.C. 20007. OR CALL:
800-821-7700 toll-free. Ask for OPTIONS.

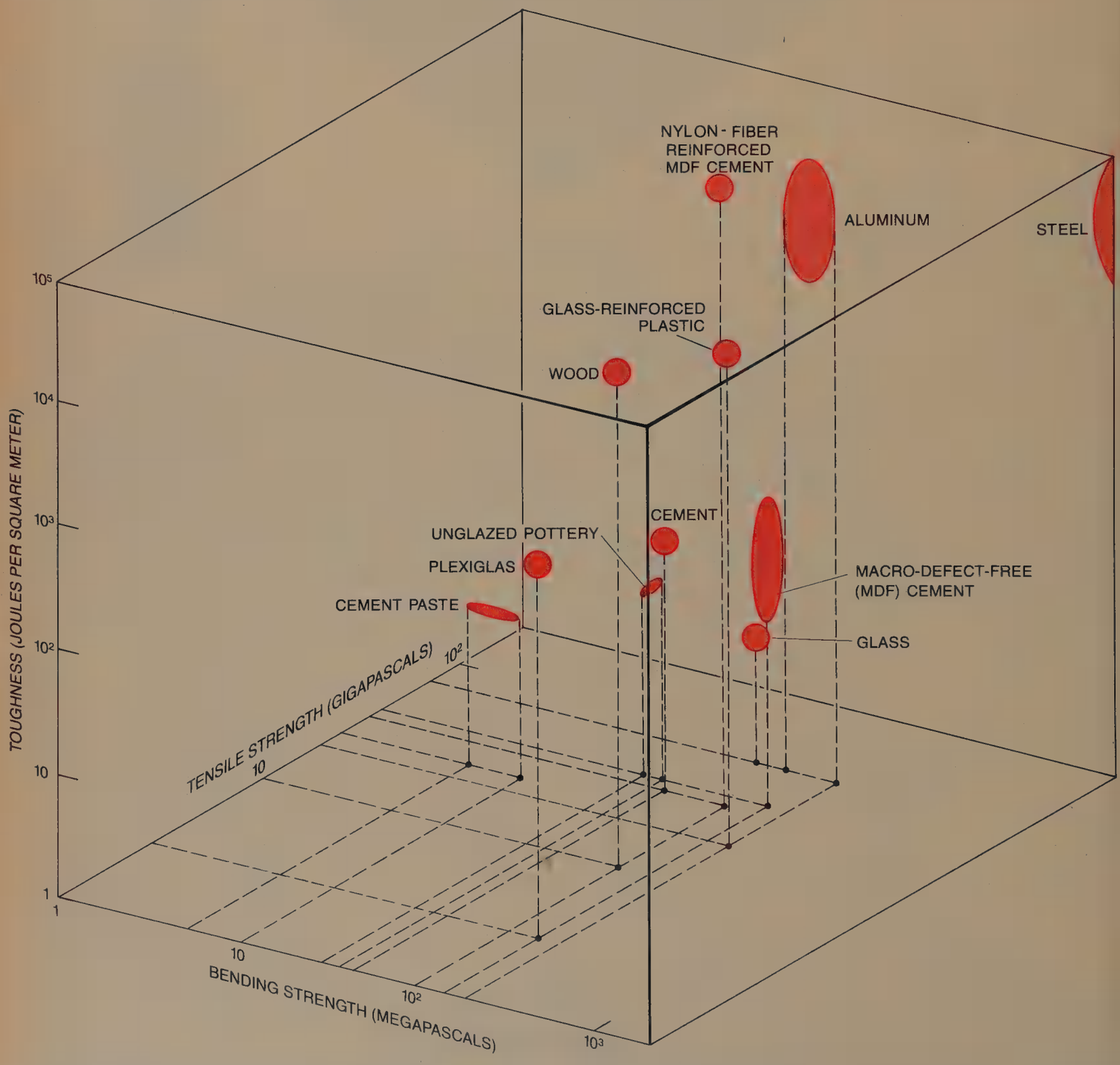
© 1983, Student Loan Marketing Association

pores remaining in the cement after it has set. The strength increases as the porosity is reduced. The strength of a typical cement is actually rather low: its flexural strength is usually less than five megapascals and its tensile strength is only seven times higher. That is why cement and concrete are employed in construction only under compressive loads and why steel reinforcement is added when the material must bear a tensile load.

Suppose the relation between porosity and strength is extrapolated to predict the strength of a cement that has no pores. The result is a flexural strength of about 20 megapascals. The flexural strength of aluminum alloys is some five times as great, and the flexural strength of steel is 50 to 100 times as great. Even if the cement were reinforced with asbestos or glass fibers, one would predict a flexural strength of no more than 40 megapascals.

The extrapolation to zero porosity, however, is now known to be invalid. The simplest theory to explain the fracture of a brittle solid was developed in 1920 by Alan A. Griffith, who was then at the Royal Aircraft Establishment at Farnborough in England. He proposed that the tensile strength of a brittle solid is determined by the size of the largest flaw in the material. If the theory is correct, the strength of a cement should increase dramatically as the diameter of its largest pores is reduced. Conversely, the strength of the cement should depend only weakly on the number of pores or their volume.

That turns out to be the case. Anthony James Howard, Kevin Kendall and one of us (Birchall), working at the Imperial Chemical Industries (ICI) Research



THREE PROPERTIES of material its stiffness (resistance to bending), its tensile strength (resistance to pulling) and its fracture toughness (resistance to impact). Here the properties are plotted for various organic and inorganic materials. The scales are logarithmic. Organic materials such as wood rate high in toughness but low in stiffness. Inorganic materials such as glass are essentially the opposite. MDF cement rates almost as high as aluminum in all three properties. Stiffness and tensile strength are in pascals; one megapascal equals 145 pounds per square inch. Fracture toughness is the energy (in joules) that opens fractures adding up to one square meter of surface.

SCIENCE/SCOPE

An Advanced Medium-Range Air-to-Air Missile has intercepted a drone target, showing its ability to find low-flying targets amid high clutter caused by the missile's radar returns reflecting from the ground. The prototype AMRAAM was fired from an F-15 fighter from an altitude of 16,000 feet and a range of about 13 miles. The remotely controlled target flew toward the F-15 only 400 feet above the ground and used electronic countermeasures in an effort to jam the missile's seeker. Hughes Aircraft Company, AMRAAM's designer, is producing the missile under a full-scale development contract for the U.S. Air Force and Navy.

The infrared-guided Maverick missile has proven its effectiveness against many kinds of targets, scoring 20 direct hits in 26 launches in evaluation tests. The IR Maverick adds precision night attack capabilities to the U.S. Air Force arsenal of air-to-surface weapons. In addition to night and day capability, its seeker sees through battlefield haze and smoke. The 20 direct hits were scored against moving tanks, a hangarette, radar vans, idling tanks, a simulated large building, a patrol boat, and a simulated fuel dump. Eleven hits came at night. Weather conditions and terrain varied from humid subtropics to desert to cold snowy plains. The misses involved minor hardware or software problems that have been corrected. Hughes has begun low-rate pilot production of 200 missiles.

An infrared sensor no larger than a collar button will be the heart of a seeker designed to improve the ability of future short-range air-to-air missiles to find and track low-flying targets. The sensor is a focal plane array, a tiny hybrid sandwich with infrared detectors on one side, each attached to a silicon signal-processing chip. It senses temperature differences in a scene and creates sharp TV-like pictures. The array is better than existing sensors because it is far more sensitive, it reduces confusion sometimes caused by ground clutter or background heat, and it can see a target rather than home on jet exhaust. Hughes is developing a chip with 4,000 sensors for the U.S. Navy. It will form one quadrant of a larger hybrid containing 16,000 detectors.

A new digitally compensated quartz crystal oscillator is designed for portable or remotely located equipment requiring low power and high stability. The Hughes oscillators represent the first use of digital techniques to compensate for frequency drift with temperature in hybrid oscillators. They use CMOS technology for minimal power dissipation, and may be customized to specific needs.

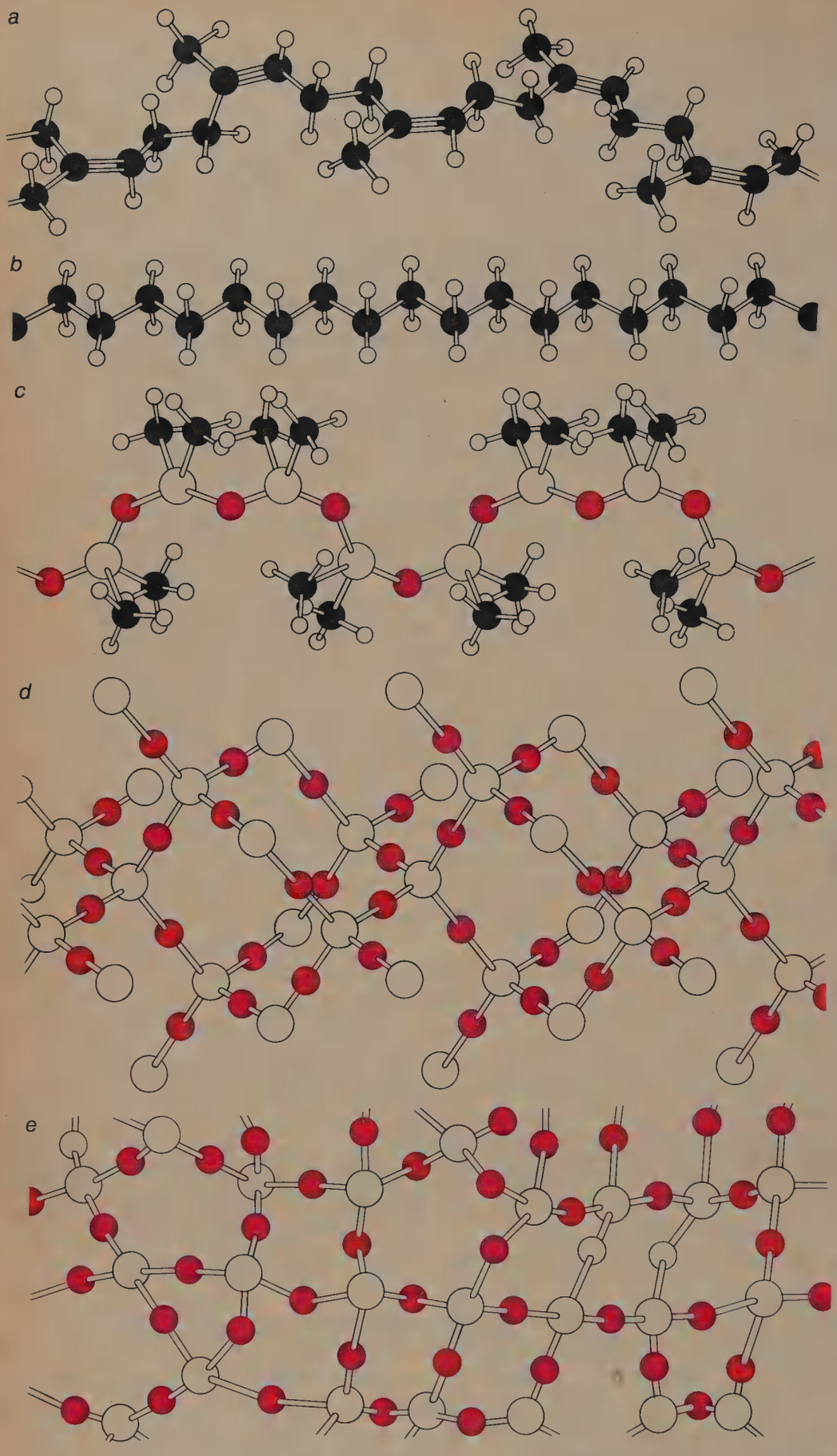
An ultramodern facility spanning 1.75 million square feet is the showcase where outstanding Hughes engineering combines advanced manufacturing techniques and production processes. Our complex is complete, so we're looking for experienced and graduating engineers to work on such programs as: infrared thermal imaging systems, laser rangefinders and designators, and missile launching and guidance systems. Send your resume to Hughes Electro-Optical and Data Systems Group, Professional Employment, Dept. SE, P.O. Box 913, Bldg. E9, M.S. W101, El Segundo, CA 90245. Equal opportunity employer.

Creating a new world with electronics

HUGHES

HUGHES AIRCRAFT COMPANY

For more information please write:
P.O. Box 11803, Los Angeles, CA 90291



POLYMERS, long molecules made up of repeating chemical units, include a variety of useful materials, both organic and inorganic. Natural rubber (*a*), an organic polymer, tends to stretch when it is pulled. The addition of sulfur atoms makes it resistant to stretching by forming bridges between its hydrocarbon chains. Polyethylene (*b*) is a synthetic organic polymer. It was introduced in the 1930's and contributed to the growth of the plastics industry. Silicone (*c*) is synthetic and inorganic; its backbone consists not of carbon but of silicon and oxygen in alternation. It represents the only commercially successful synthesis of an inorganic polymer, yet such polymers are common in nature. The mineral cristobalite (*d*) is a polymer of silicon and oxygen, as is silicate glass (*e*).

Laboratories at Runcorn in England, have now shown that cement obeys Griffith's theory. Moreover, we have developed ways in which the maximum diameter of the pores in cement can be reduced from the usual millimeter or so to no more than a few micrometers. In essence the pores are kneaded out. In order to make the kneading thorough the cement grains are graded in size and a small quantity of a polymer is added as an aid to flow.

The new material has become known as macro-defect-free (MDF) cement. Tested in bending it shows a strength of more than 150 megapascals and in that respect is comparable to aluminum. Indeed, it is entirely possible to use cement to fabricate a spring.

The fracture toughness of MDF cement is also notable. Most common manmade inorganic materials have a fracture toughness of less than 10^2 joules per square meter of new surface generated by fracturing. In contrast the fracture toughness of wood is 10^4 joules per square meter across the grain. (It drops to 10^2 joules per square meter with the grain.) The mineral jade has a fracture toughness of 10^3 joules per square meter; that is why it can be carved delicately to shape. The fracture toughness of MDF cement is also about 10^3; on a conventional lathe, therefore, a block of it can be hollowed into a tube, a remarkable performance for an inorganic material.

Still, no cement is highly resistant to impact until it is reinforced with fibers. The technology of doing so has recently matured. In fact, strips of ordinary cement can now be made pliable. They can be bent like strips of metal. The bending does make cracks in the cement, but the cracks are quite narrow and the reinforcing fibers hold the cement together. Since the cement is made at low temperature, the fibers can be inexpensive organic ones that have a low melting point.

The new high-strength cement has the appearance of a ceramic; indeed, it would make good cups and saucers. It also has many of the characteristics of a ceramic. It is strong, hard, brittle and stiff, and it is formed largely by covalent chemical bonding. It is not, however, a refractory: a material that keeps its integrity at high temperature. The reason is that it contains water in chemical combination with its calcium silicates. At high temperature the water is driven off, leaving behind a much weaker solid. It would obviously be of value to devise methods of manufacturing refractory materials at low temperature.

Two approaches are promising. We noted above that the silicate rocks and minerals in the earth's crust are polymers in which alternating oxygen atoms

DATSUN 200-SX

IT'S GOT PERFORMANCE DOWN TO A SCIENCE.

The sleek Datsun 200-SX is a scientific paradox. Although it burns gas "faster" than conventionally powered cars, the 200-SX gets even better gas-mileage ratings this year than last. And last year's ratings were already remarkably high for a car that won two dozen races in just nine months.

How? The 2.2-liter Nissan NAPS-Z engine. A microprocessor continuously adjusts the fuel/air mixture injected into the engine's crossflow hemihead combustion chambers and selects which of each cylinder's twin sparkplugs to fire, a unique concept in engine design.

Combustion in this "fast-burn" system is completed within 90° of crankshaft rotation. Ordinary engines don't complete combustion until 170°, wasting both power and fuel. The Nissan NAPS-Z saves both by burning gas "faster."

Most important, you can have this scientific advance without compromising looks or luxury features. This unique power plant delivers its high performance wrapped in a sculptured body that turns science into art. See and test drive the 200-SX at your Datsun dealer now.

38 EPA EST. HWY. (27) EPA EST. MPG

Use EPA estimated mpg for comparison, with standard 5-speed. Actual mpg may differ depending on speed, trip length and weather. Highway mpg will probably be less.

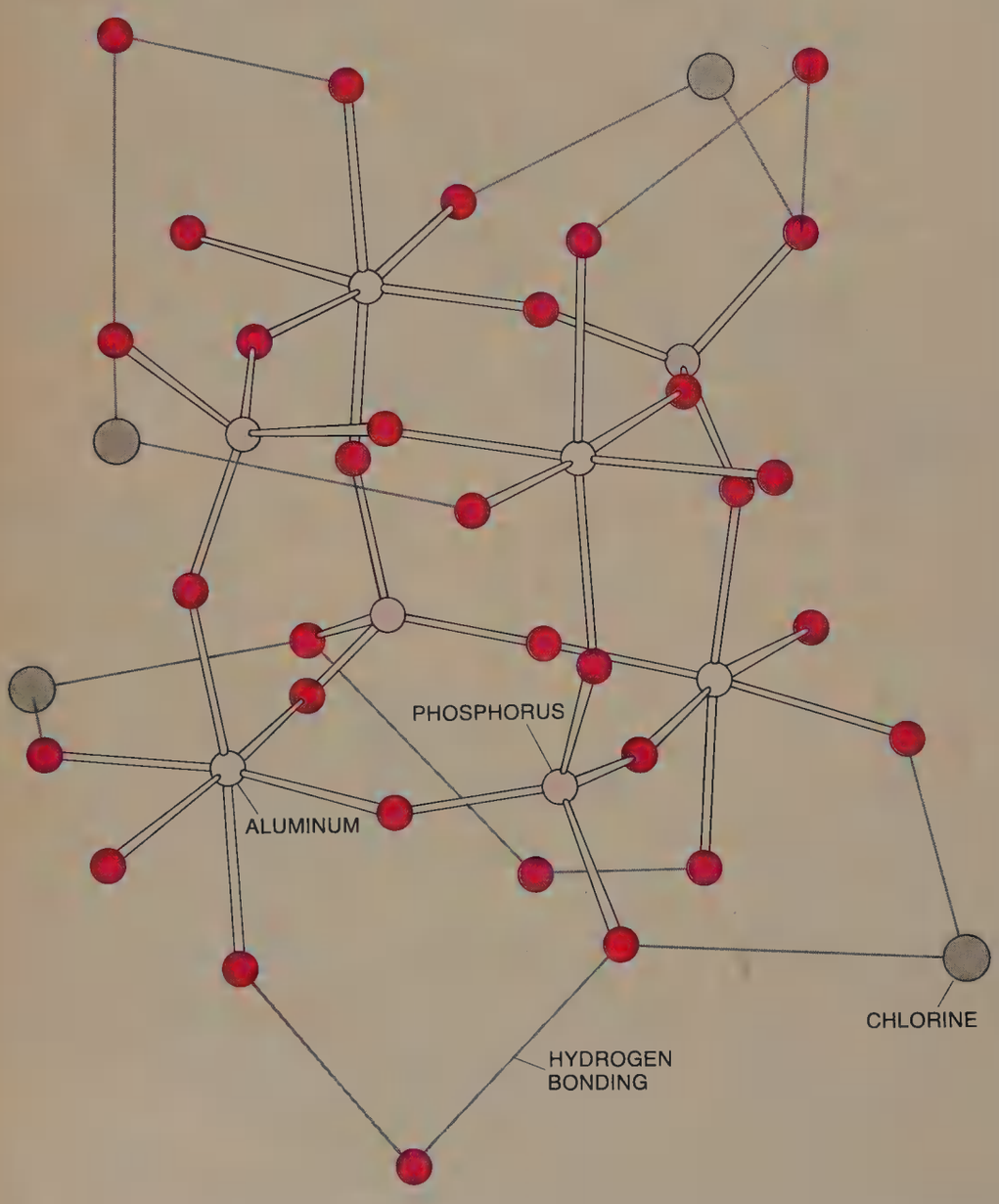
NISSAN
WE ARE DRIVEN
DATSUN

NISSAN NISSAN MOTOR CORPORATION IN USA

and silicon atoms form chains and rings. In some of these minerals the chains and rings lie in silicate sheets some tens of angstroms thick. The sheets are separated from each other by layers of cations (positive ions) such as magnesium. One such mineral is vermiculite. George F. Walker of the Division of Applied Mineralogy of the Australian Commonwealth Scientific and Industrial Research Organization has discovered that the silicate sheets can be separated by removing or replacing the cations and then applying a shearing force. The result is a multitude of shards (fragments of the silicate sheets) thickly suspended in water. The suspension can be dried to make a strong, flexible, translucent and refractory film. Alternatively, the suspension can be whisked and then allowed to dry to make a refractory inorganic foam much like the organic polymer polystyrene. Here, then, is the complete re-creation of a natural mineral.

The second approach to the manufacture of refractory inorganic materials at low temperature is emerging from the study of solids such as glassy aluminum phosphate ($AlPO_4$). Aluminum phosphate is closely akin to silica, the archetypical inorganic refractory ceramic, and in fact it displays similar crystal forms. It is virtually impossible to produce glassy aluminum phosphate merely by melting the compound itself. The compound decomposes before it melts. On the other hand, a crystalline precipitate with the composition $AlPO_4 \cdot HCl \cdot (C_2H_5OH)_4$ can be prepared by reacting aluminum trichloride ($AlCl_3$) and phosphoric acid (H_3PO_4) in ethyl alcohol (C_2H_5OH) at temperatures lower than zero degrees C.

The structure of the precipitate is curious. It consists of aluminum, oxygen and phosphorus atoms in a cubic arrangement, but the cubes are prevented from linking by intervening molecules of ethyl alcohol. On gentle heating to about 100 degrees C. the alcohol is lost and the cubes do link. They form a three-dimensional lattice of aluminum, oxygen and phosphorus that is glassy and is refractory up to 1,600 degrees C., the temperature at which aluminum phosphate begins to decompose. The glass is inert and insoluble, but the precipitate from which the glass emerges is soluble both in water and in organic solvents. The precipitate can thus be dissolved and warmed. The process yields glassy coatings. Moreover, the precipitate can be used to bind together particles of compounds such as alumina (Al_2O_3) and so can serve the creation of a ceramic at a temperature of about 100 degrees C.

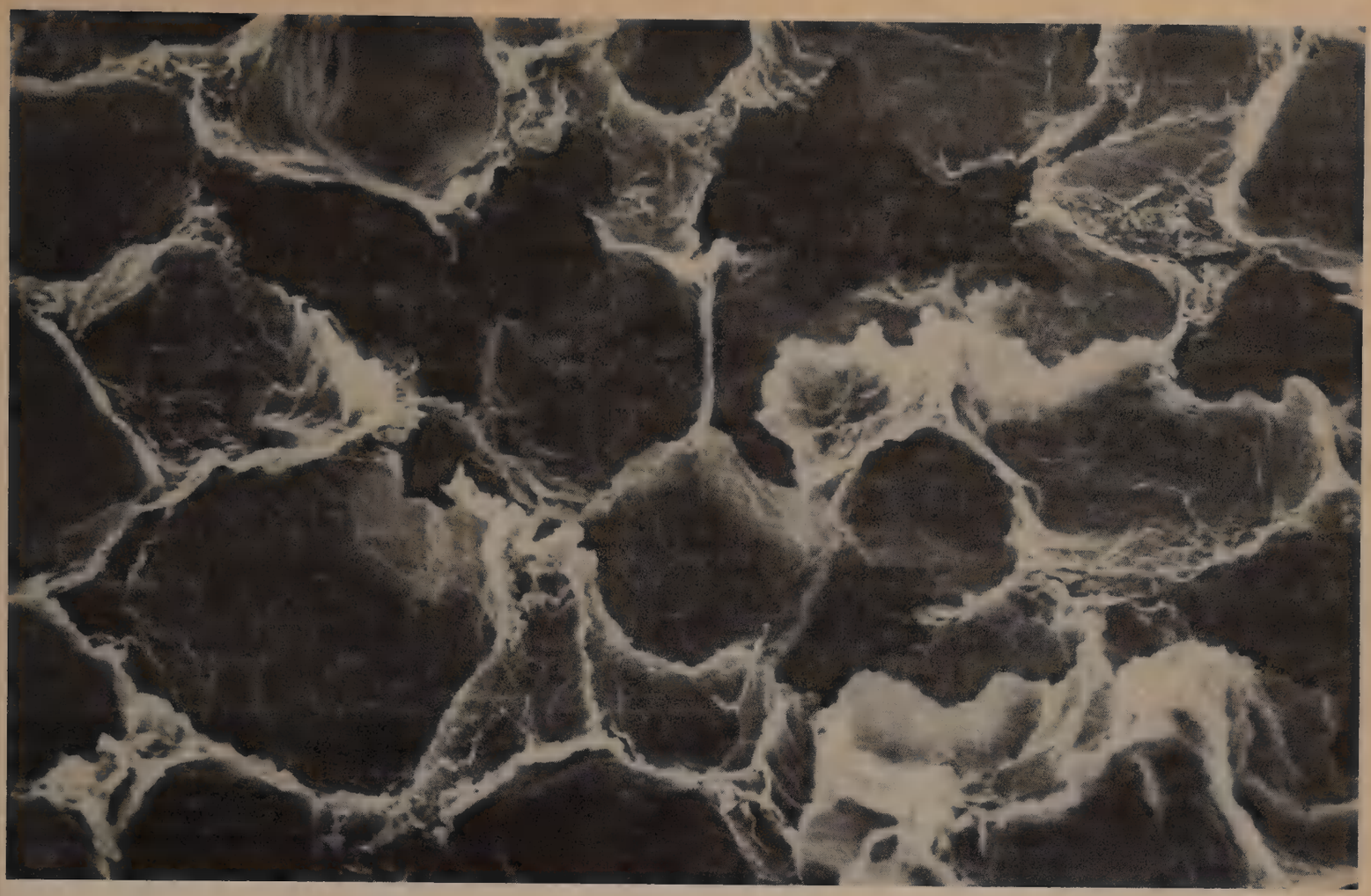


GLASSY ALUMINUM PHOSPHATE ($AlPO_4$) emerges from the reaction of aluminum trichloride ($AlCl_3$) with phosphoric acid (H_3PO_4) in ethyl alcohol (C_2H_5OH). The reaction itself yields the intermediate material shown here. Aluminum and phosphorus atoms form the vertexes of an approximate cube; oxygen atoms lie along its edges. The cubes are separated from one another by ethyl alcohol molecules. At a temperature of approximately 100 degrees Celsius the alcohol is lost and the cubes coalesce into a glass. The structure of the intermediate material was determined by John E. Cassidy and his colleagues at Imperial Chemical Industries.

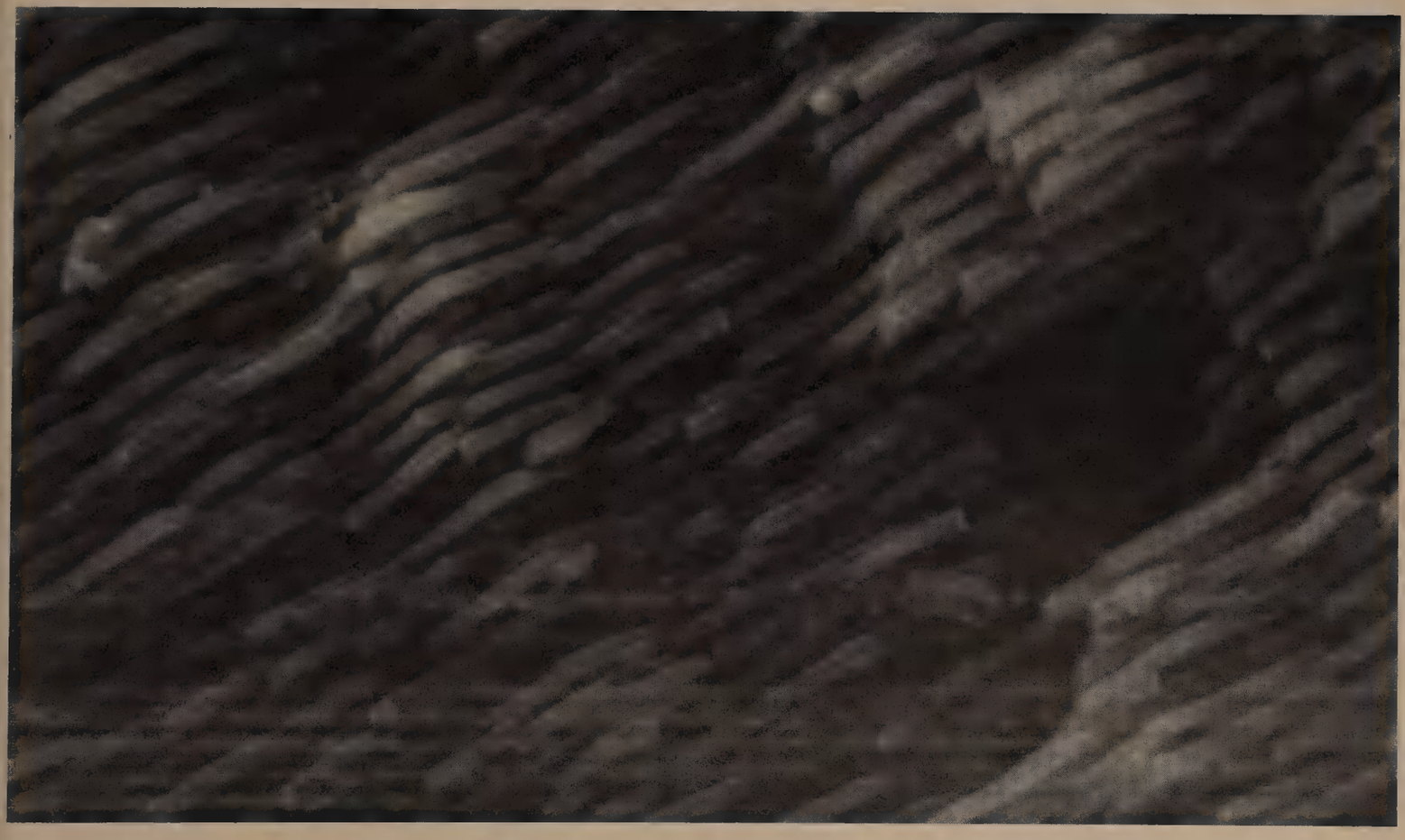
In related techniques developed by B. E. Yoldas of the Westinghouse Research Laboratories and by Masayuki Yamane and Tesuo Sakaino of the Tokyo Institute of Technology glasses are made at low temperature from alkoxides: compounds that emerge from the reaction between an alcohol and an inorganic acid. Methyl alcohol (CH_3OH), for example, reacts with silicic acid ($Si(OH)_4$) to make silicon alkoxide ($Si(OCH_3)_4$) plus water. If the alkoxide is reacted with water, it is transformed into an oxide and an alcohol. The alcohol can be removed by gentle heating, leaving a glassy oxide behind. In principle the glass could be reinforced with organic fibers that have a low melting point, a strategy that is not possible in glass made conventionally at high temperature.

Like the cement described above, the glass produced at low temperature is porous. Here again, however, the pores are small, and so they have little effect on many physical properties. In making glass one often aims to control not only the mechanical properties of the material but also the optical properties. In particular the transparency of a glass requires that no pores be as large as a wavelength of the light the glass is intended to transmit. Many of the glasses produced by Yoldas are splendidly transparent; they contain no pores larger in diameter than 100 angstroms.

In the ways we have described, inorganic chemistry is preparing to contribute to the conservation of energy and of hydrocarbons by manipulating inorganic chemical compounds at low temperature. This manipulation is being aided by advances in the understanding of material properties such as porosity. Interestingly, the importance of the control of porosity had not emerged from all the previous work on high-temperature methods of fabrication. It may be too early to speak of a new Neolithic age, but its beginnings are clearly with us.



FOAM MADE FROM VERMICULITE is a newly devised inorganic material that can supersede organic polymers such as polystyrene plastic. In vermiculite, silicate sheets are interleaved by layers of magnesium ions. The ions are removed and the sheets are sheared. The resulting slurry is whisked and dried to produce the foam. This scanning electron micrograph was made at an enlargement of 75 diameters.



ABALONE SHELL, seen in a scanning electron micrograph, is almost entirely chalk (calcium carbonate, $CaCO_3$), yet it is stiffer than aluminum and as tough as Plexiglas. It confirms that inorganic materials can be designed to have new and useful properties. The shell consists of calcium carbonate plates that are .2 micrometer thick and are separated from each other by a thin layer of protein.

Computer-Intensive Methods in Statistics

They replace standard assumptions about data with massive calculations. One method, the "bootstrap," has revised many previous estimates of the reliability of scientific inferences

by Persi Diaconis and Bradley Efron

Most statistical methods in common use today were developed between 1800 and 1930, when computation was slow and expensive. Now computation is fast and cheap; the difference is measured in multiples of a million. In the past few years there has been a surge in the development of new statistical theories and methods that take advantage of the high-speed digital computer. The new methods are fantastic computational spendthrifts; they can easily expend a million arithmetic operations on the analysis of 15 data points. The payoff for such intensive computation is freedom from two limiting factors that have dominated statistical theory since its beginnings: the assumption that the data conform to a bell-shaped curve and the need to focus on statistical measures whose theoretical properties can be analyzed mathematically.

These developments have profound implications throughout science, because statistical theory addresses a grand question: How is one to learn what is true? Suppose 15 measurements of some quantity yield 15 moderately different values. What is the best estimate of the true value? The methods of statistics can answer such a question and can even give a quantitative indication of the estimate's reliability. Because empirical observations are almost always prone to error, conclusions in the sciences (and in many other fields) must often be based on statistical measures of truth. As a result any development that makes statistical inferences more accurate or more versatile can be expected to have broad consequences.

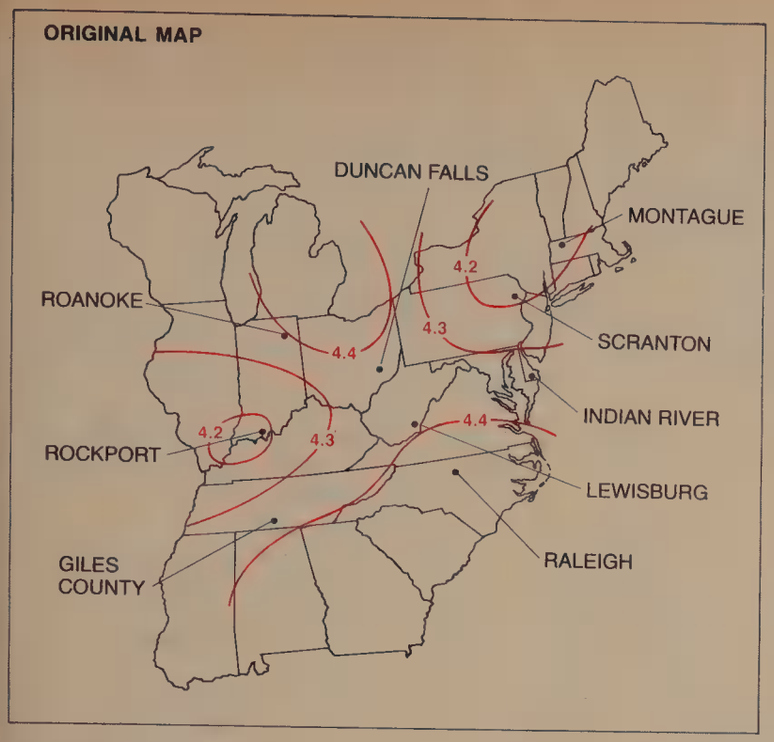
The two advantages of the new methods are best appreciated by comparing them with older ones. First, in older methods it was generally necessary to make certain unverifiable assumptions about the data before statistical analysis could proceed. The assumptions often involved the bell-shaped curve, which is also called the normal or Gaussian distribution, after the German mathematician Carl Friedrich Gauss. When the Gaussian distribution is employed, it is assumed that random fluctuations, or errors, in the experimentally observed values of some quantity are scattered symmetrically about the true value of the quantity. Moreover, it is assumed that the greater the error between the experimental value and the true value is, the less likely it is that the experimental value will be observed. Experience has shown that Gaussian theory works quite well even when the Gaussian distribution is only roughly approximated by the data, which is why statisticians can give reliable predictions even without computers. For sets of data that do not satisfy the Gaussian assumptions, however, the results of statistical methods based on such assumptions are obviously less reliable. Computer-intensive methods can solve most problems without assuming that the data have a Gaussian distribution.

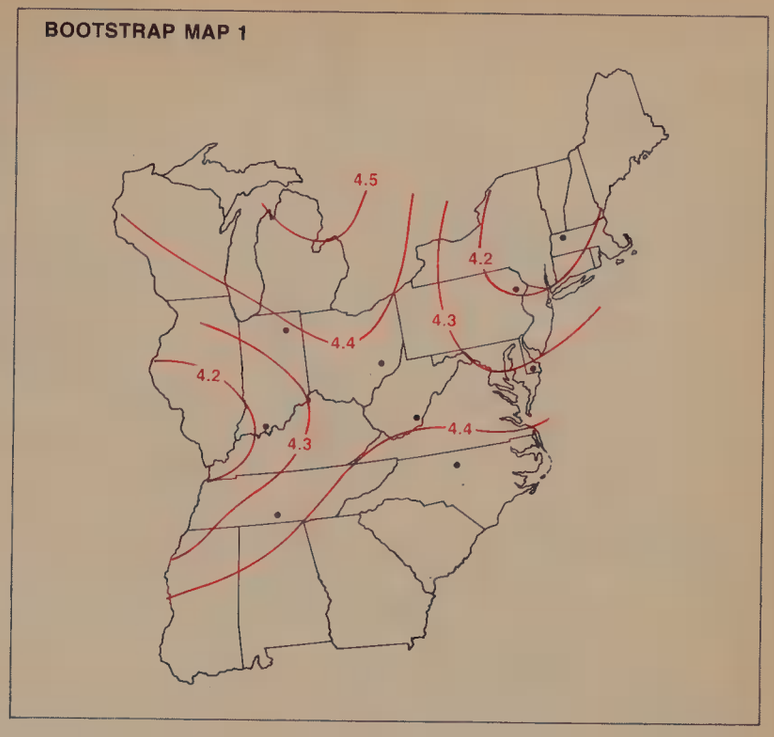
Freedom from the reliance on Gaussian assumptions is a signal development in statistics, but the second advantage of the new techniques is even more liberating. In older practice the arithmetic operations associated with statistical analysis had to be done by hand or with the aid of a desk calculator. Such calculations can often be simplified immensely if the formulas on which the calculations are based have a concise analytical form. Hence statistical theory tended to focus on only a few properties of a statistical sample, such as the mean, the standard deviation and the correlation coefficient, that can easily be manipulated analytically. Many other properties of a sample, however, are of interest to the statistician but are beyond the reach of exact mathematical analysis. The new computer-based methods make it possible to explore such properties numerically, even though their exact analysis is currently impossible. Thus the new methods free the statistician to attack more complicated problems, exploiting a wider array of statistical tools.

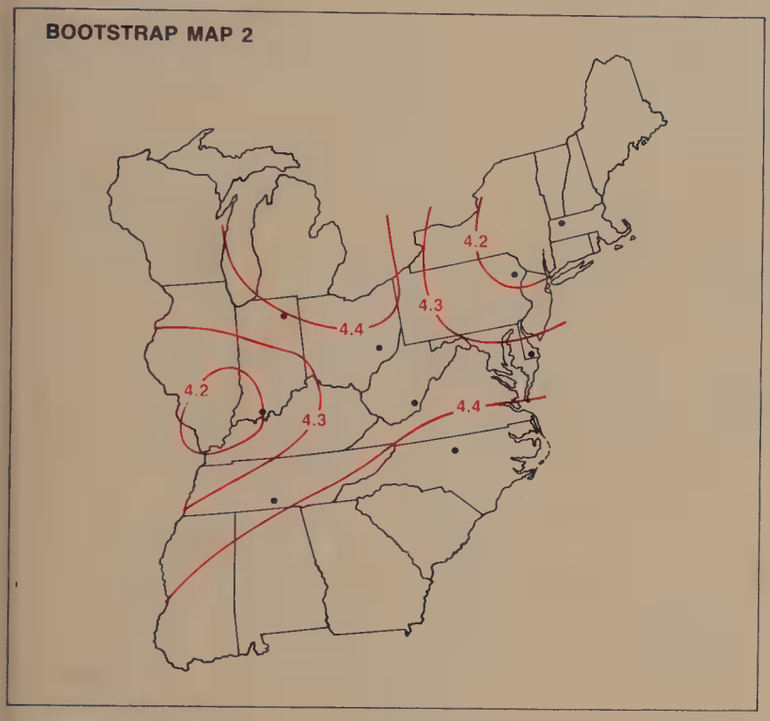
To illustrate how the computer has been applied to statistical inference we have chosen a problem in which only 15 data points appear. We shall apply a method called the bootstrap, invented by one of us (Efron) in 1977, which is quite simple to describe but is so dependent on the computer that it would have been unthinkable 30 years ago.

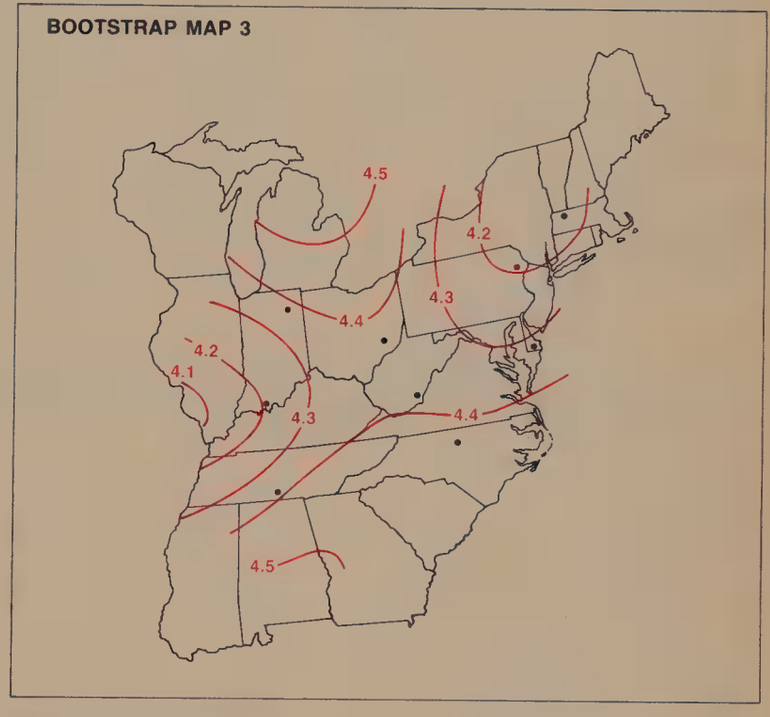
Consider a group of 15 law schools for which two overall characteristics of each entering freshman class are measured: the average undergraduate grade-point average (GPA) and the average

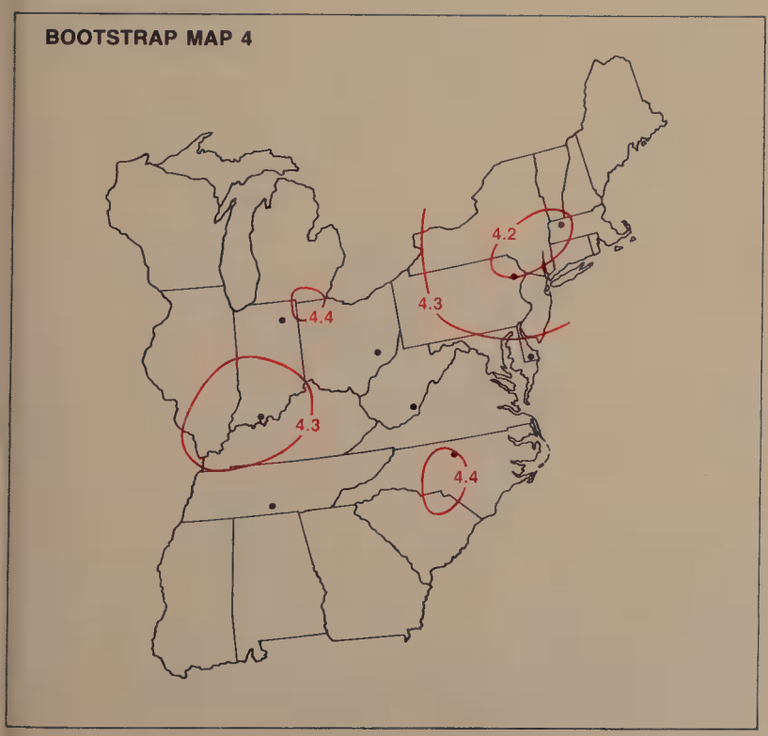
LARGE VARIABILITY of contour lines on a map is revealed by the statistical method called the bootstrap; the method requires so many numerical calculations that it is feasible only with the aid of a computer. The map at the upper left was constructed from 2,000 measurements of the acidity, or *p*H value, of every rainfall recorded at nine weather stations over a period of two years. (The lower the value of the *p*H, the greater the acidity.) The contours were drawn according to a procedure that can be proved optimal under certain conditions. Nevertheless, the 2,000 data points are subject to considerable random variability: contours based on another sample of 2,000 measurements for the same region and time period might have looked quite different. The bootstrap, which was invented by one of the authors (Efron), can estimate from the single set of 2,000 data points the amount of variability the contours would show if many sets of 2,000 data points could be compared. The results of five bootstrap calculations, done with the aid of a computer by Barry P. Eynon and Paul Switzer of Stanford University, are shown in the other five maps. The variability of the contours shows that the original map must be interpreted cautiously: corridors of low acidity on the original map can become islands on subsequent maps.

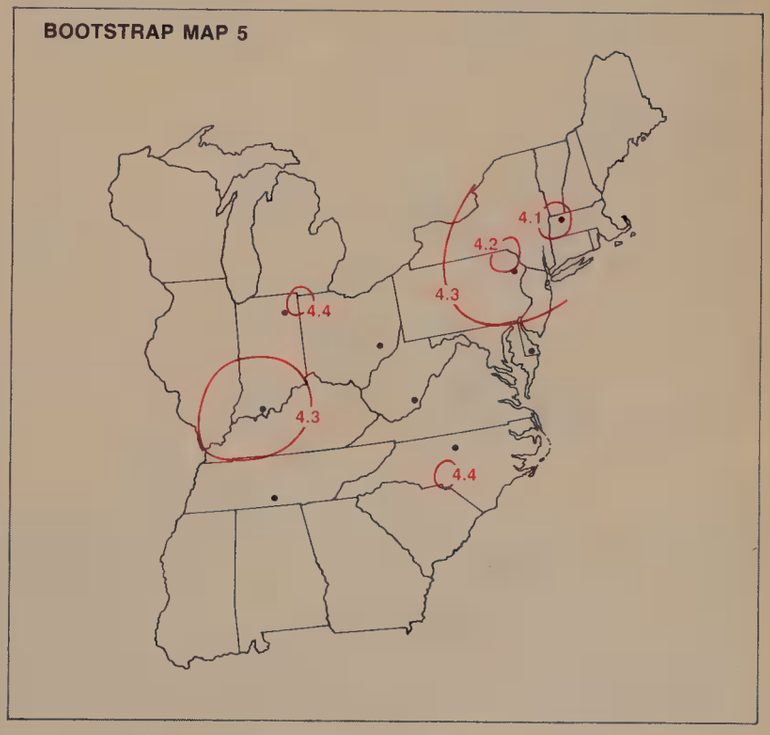

ORIGINAL MAP
DUNCAN FALLS
MONTAGUE
ROANOKE
SCRANTON
INDIAN RIVER
ROCKPORT
LEWISBURG
RALEIGH
GILES
COUNTY
4.2
4.3
4.4

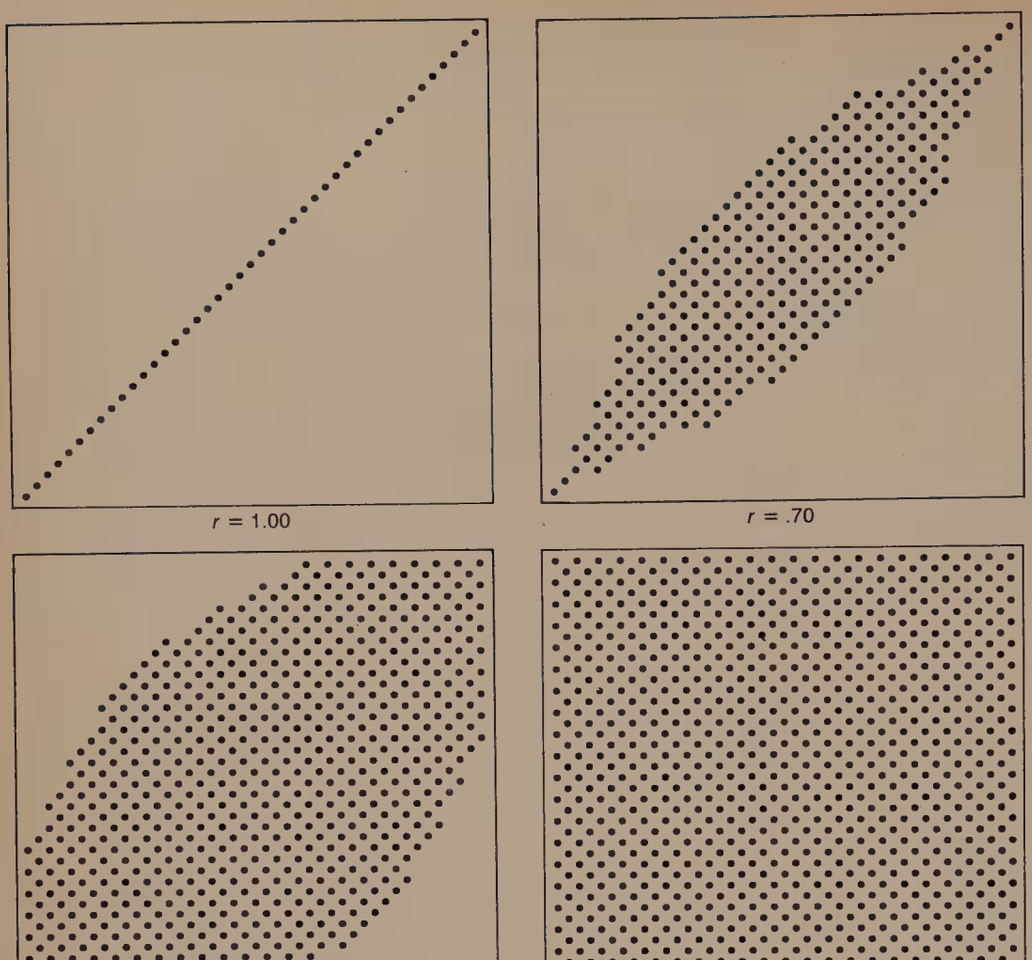

BOOTSTRAP MAP 1
4.5
4.2
4.3
4.4


BOOTSTRAP MAP 2
4.2
4.3
4.4

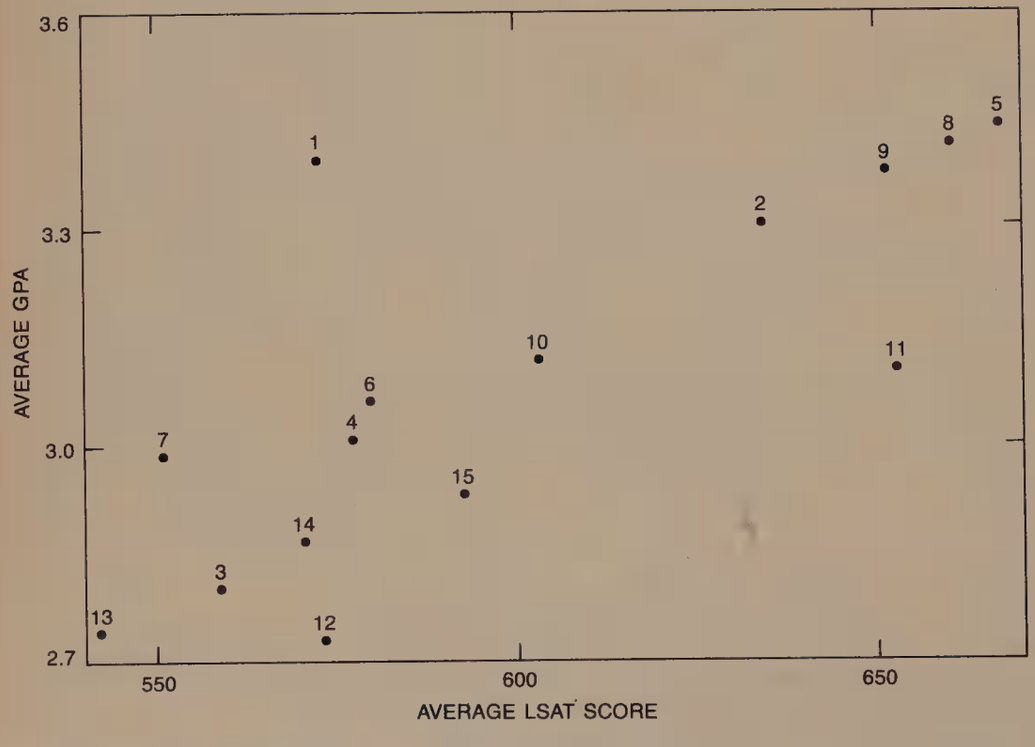

BOOTSTRAP MAP 3
4.5
4.2
4.3
4.4
4.1


BOOTSTRAP MAP 4
4.2
4.3
4.4


BOOTSTRAP MAP 5
4.1
4.2
4.3
4.4



CORRELATION COEFFICIENT is a measure of the tendency of data points plotted on a graph to cluster about a line. The coefficient is usually designated by the letter *r* and can have any value between 1 and −1. The more linear the clustering, the greater the absolute value of *r*.



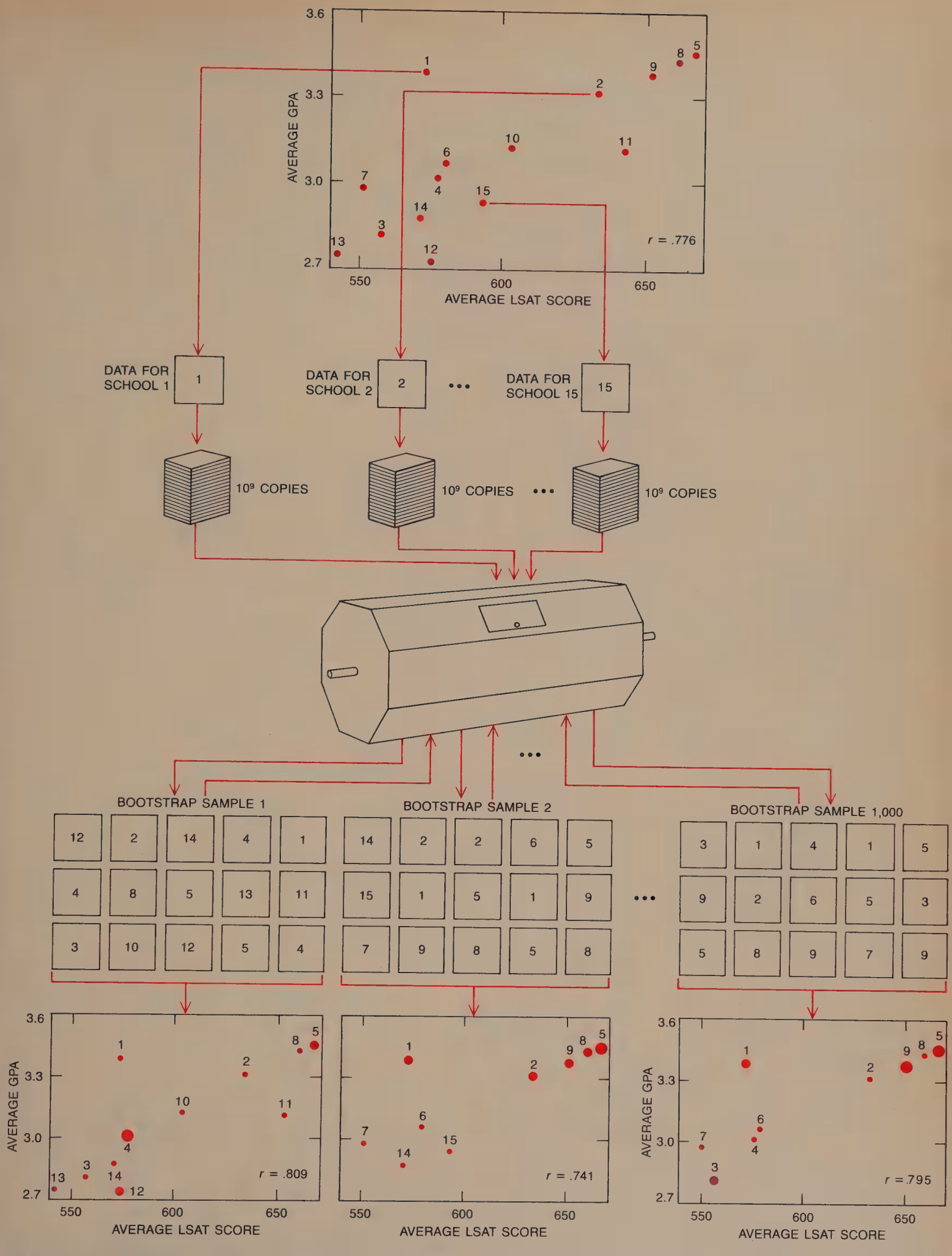
TWO MEASURES for the potential academic performance of the students in the entering classes of 15 American law schools are graphed for each school. Each point on the graph represents the undergraduate grade-point averages (GPA) and the scores on the Law School Admission Test (LSAT), averaged over all the students in one entering class. The graph shows that for the sample of 15 law schools the two measures tend to be proportional: their correlation coefficient *r* is .776. One would like to know how accurately .776 approximates the true value of *r* for all American law schools. That is, one would like to know how much, on the average, the observed value of *r* for a random sample of 15 law schools differs from the true value of *r*.

score on the Law School Admission Test (LSAT). It seems reasonable to suppose the two measures ares roughly proportional to each other: the entering classes with a high average GPA tend to have a high average score on the LSAT. It is unlikely, however, that the proportionality is exact: the entering classes of one or two law schools may show a high average GPA but a low average LSAT, whereas a few other schools may have a low average GPA but a high average LSAT. The statistician wants to know first of all how close the relation between the two measures is to proportionality. Moreover, the statistician must try to estimate the degree to which the available data justify the extrapolation of the first result to all other law schools. In short, how confident can one be that the sample of 15 law schools gives an accurate picture of the population of law schools as a whole?

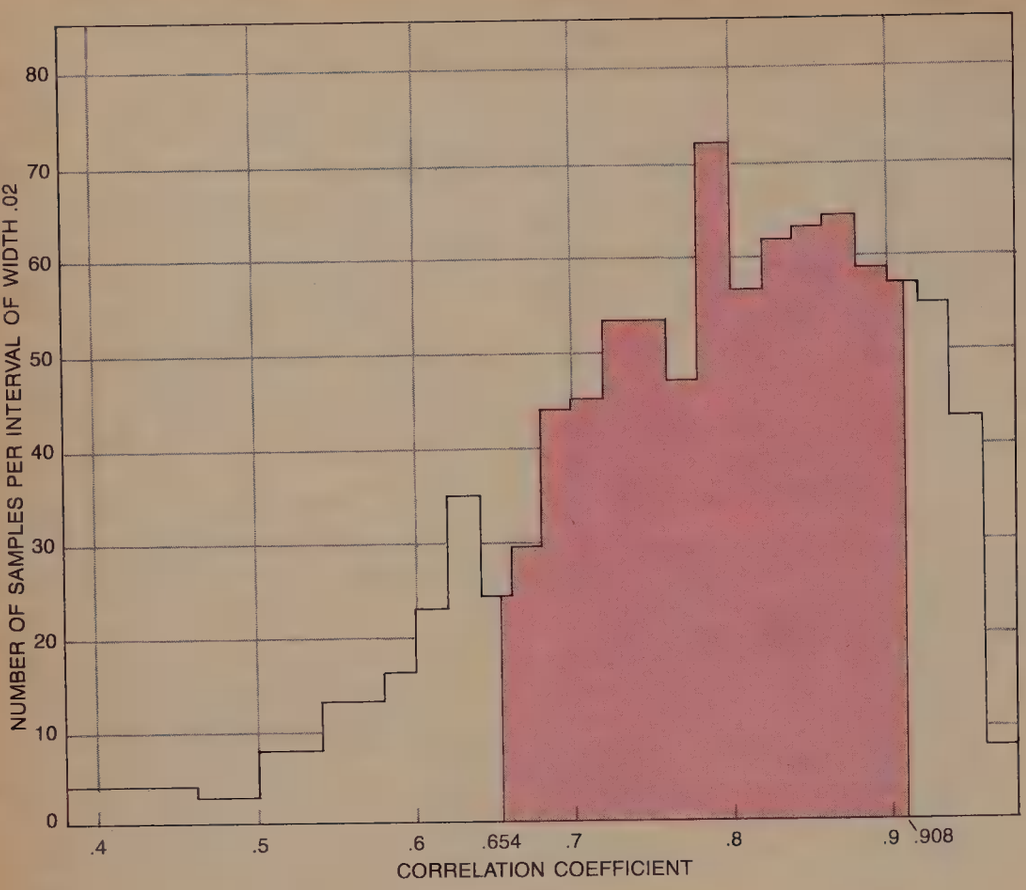
The standard measure of the tendency toward proportionality between two variables such as GPA and LSAT is the correlation coefficient; it is usually designated by the letter *r*. Suppose the data for the law schools are plotted on a graph where the vertical axis represents GPA and the horizontal axis represents LSAT. The correlation coefficient is a measure of the degree to which the points on such a graph tend to cluster along a line. The value of *r* is 0 if the points are scattered at random and gets increasingly close to 1 or −1 as the points tend to cluster along a line of positive or negative slope. (The slope of a line is positive if the line slopes up and to the right, and the slope is negative if the line slopes down and to the right.) The correlation between degrees Fahrenheit and degrees Celsius, for example, is 1 because the two variables are directly proportional to each other. The correlation between the height of fathers and the height of their sons is about .5. Tall fathers tend to have tall sons, but the correspondence is not exact. The correlation between daily consumption of cigarettes and life expectancy has been shown to be negative; that is, the greater the daily consumption of cigarettes, the shorter the life expectancy.

The observed correlation between GPA and LSAT for the 1973 entering classes of the 15 law schools is .776. In other words, there is a strong correlation observed between the two variables and a strong tendency for the points defined by the coordinates of each school to cluster along a line of positive slope. A straightforward mathematical procedure, which takes about five minutes with a desk calculator, was applied to determine the value of *r*: the details of the calculation are not important except that they give a well-defined value of *r* for any collection of data points.

What grounds does one have, how-



BOOTSTRAP METHOD is applied to the sample of 15 law schools shown in the illustration at the bottom of the opposite page in order to assess the accuracy of the correlation coefficient calculated for the sample. The data for each law school are copied perhaps a billion times and all 15 billion copies are thoroughly shuffled. Artificial samples of 15 law schools, called bootstrap samples, are created by selecting sets of 15 data points at random from the 15 billion copies. The value of r is then calculated for each of the bootstrap samples. Although it is simple in concept, the application of the bootstrap requires so many calculations that it is not feasible without a computer.



FREQUENCY DISTRIBUTION of the correlation coefficient *r* is plotted for 1,000 bootstrap samples. A widely accepted measure of the accuracy of a statistical estimator such as *r* is the width of the strip under the central part of its frequency distribution whose area is 68 percent of the area under the entire distribution. The central strip for the bootstrap distribution is shaded in color; its width is .254. Half of the width of the interval, .127, is a good estimate of the average amount by which the observed value of *r* for a sample differs from the true value of *r*.

ever, for believing the true value of *r* is close to .776 for all law schools? After all, the sample could be highly atypical of law schools in general. The law of large numbers guarantees that in large samples the statistical estimate of *r* calculated for the sample is very likely to approach the true value of *r* for the entire population. A sample of only 15 law schools, however, is not a large sample. Hence some measure is needed that can assess the statistical accuracy of the value of *r* given by the sample, namely .776. The bootstrap is intended to provide such a measure.

To understand what statistical accuracy means for an estimate such as *r*, suppose data were available for additional sets of 15 law schools, different from the sets already sampled. For each set of 15 law schools the value of *r* could be calculated, and the amount of variation in the values of *r* for many samples could thereby be described. For example, if 99 percent of the values of *r* calculated for the hypothetical samples were between .775 and .777, one would assign high accuracy to the estimate .776. On the other hand, if the values of *r* were spread out evenly from -1 to 1, the estimate of *r* given by the original sample would have no accuracy and would therefore be useless. In other words, the statistical accuracy of an estimated value of *r* depends on the width of the interval bracketing the estimated value that is associated with a certain percentage of all the samples. Unfortunately the data needed to calculate the value of *r* for many different samples are generally lacking. Thus, because the law school example is intended to reflect the conditions of real statistical practice, we shall assume for the moment that the only data available are those for the original sample of 15 law schools. Indeed, if more data were available, they could be employed to give a better estimate for the value of *r* than .776.

The bootstrap procedure is a means of estimating the statistical accuracy of *r* from the data in a single sample. The idea is to mimic the process of selecting many samples of size 15 in order to find the probability that the values of their correlation coefficients fall within various intervals. The samples are generated from the data in the original sample. The name bootstrap, which is derived from the old saying about pulling yourself up by your own bootstraps, reflects the fact that the one available sample gives rise to many others.

In effect, the bootstrap samples are generated as follows. The data for the first school are copied an enormous number of times, say a billion, and the data for each of the other 14 schools are copied an equal number of times. The resulting 15 billion copies are thoroughly mixed. Samples of size 15 are then selected at random and the correlation coefficient is calculated for each sample. On a computer the steps of copying, mixing and selecting new sets of data are all carried out by a procedure that is much faster but mathematically equivalent: the computer assigns a number to each law school and then generates the samples by matching a string of random numbers to the numbers that correspond to the law schools.

The samples generated in this way are called bootstrap samples. The distribution of the correlation coefficients for the bootstrap samples can be treated as if it were a distribution constructed from real samples: it gives an estimate of the statistical accuracy of the value of *r* that was calculated for the original sample. We generated 1,000 bootstrap samples from the data for the 15 law schools in our sample. Of the 1,000 samples there were 680, or 68 percent, whose correlation coefficients were between .654 and .908. The width of this interval, .254, is the bootstrap measure of accuracy of the value of *r* for the sample. Half of the width of the interval, .127, can be interpreted as the bootstrap estimate of the average amount by which the observed value of *r* for a random sample of size 15 differs from the true value of *r*.

It is worth noting that the statistical accuracy cannot be defined simply as the accuracy of an individual estimate such as .776, that is, the difference between the estimate and the true value of *r*. In a real problem this difference can never be known; if it were known, the problem would vanish, because one could subtract the difference from the estimate and so obtain the true value exactly. Instead the statistical accuracy refers, as we have indicated, to the average magnitude of the deviation of the estimate from the true value.

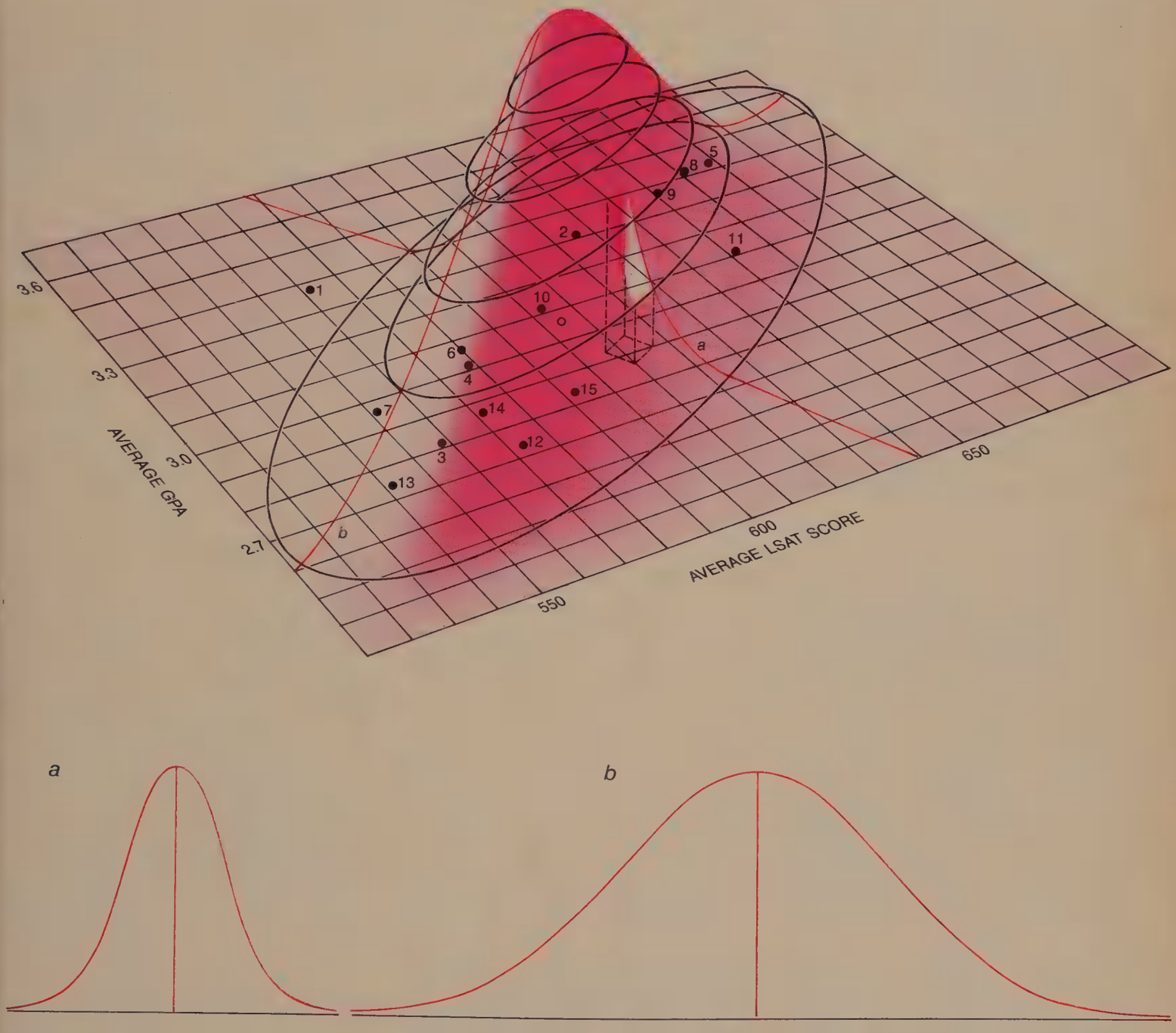
If the results of the bootstrap distribution can be taken as a measure of the statistical accuracy of the original estimate, then .776 is a rough estimate but not entirely worthless. The true correlation coefficient, that is, the value of *r* for the population of law schools as a whole, could well be .6 or .9, but it is almost certainly not zero. Our theoretical work shows that the bootstrap measure of statistical accuracy is dependable in a wide variety of situations.

We can now abandon our self-imposed ignorance because in the law school example the accuracy of the estimated correlation coefficient can be tested directly. Indeed, we chose the example because all the data for average GPA and average LSAT scores of American law school students in 1973

have already been gathered. There were 82 American law schools in 1973, and the correlation between GPA and LSAT for all the schools was .761. (Hence .761 is the true value of r we referred to above, a quantity that would not be known in most situations.) More important, the true statistical accuracy of the estimate given by the original sample can be calculated, because the distribution of the values of r for many real samples of size 15 can be determined. Samples of size 15 can be chosen at random from the 82 law schools in 82^{15}, or about 5×10^{28}, equally likely ways. In principle the value of r could be calculated for each sample, and so the number of samples for which r lies within a small interval could be plotted for intervals of equal size. The resulting graph is called a frequency distribution.

In practice the frequency distribution for samples of size 15 selected from the 82 law schools can only be approximated; a computer set to work at the beginning of the big bang calculating r for each of the 82^{15} samples at a rate of a billion a second would still not have finished the task. Instead r is calculated for a large but manageable number of randomly selected samples of size 15, say a million samples.

We found that 68 percent of the correlation coefficients for a million samples were grouped between .606 and .876, an interval whose width is .270. In other words, if a sample of 15 law schools is selected at random, the probability that its correlation coefficient lies between .606 and .876 is .68. Note that the width of the interval is in good agreement with that defined by 68 percent of



BELL-SHAPED SURFACE was employed in 1915 by Sir Ronald Fisher in his method for estimating from a single sample how much the correlation coefficient varies from sample to sample. In order to make such an estimate by Fisher's method it is necessary to assume that all the data points in the sample are selected according to probabilities given by the bell-shaped surface. The surface is constructed to fit the data in the sample. In the law school example the highest point of the surface must lie directly over the point on the plane where the GPA and the LSAT points both have their overall average values (*open circle*). The slope and orientation of the surface with respect to the plane of the graph depend on how the data points are scattered. The contours of equal height on the surface are elliptical, and the cross sections are bell-shaped curves of varying width; two cross sections are shown in the lower part of the illustration. Fisher's method can be interpreted as choosing bootstrap samples from among all the points on the plane of the graph. The probability of choosing a point from within a given region on the graph is equal to the volume that lies between that region and the bell-shaped surface (*volume of "hole"*) divided by the entire volume that lies between the surface and the graph. By carrying out the bootstrap sampling with a computer only for the discrete points in the original sample the probability distribution that is given by the bell-shaped surface need not be assumed.

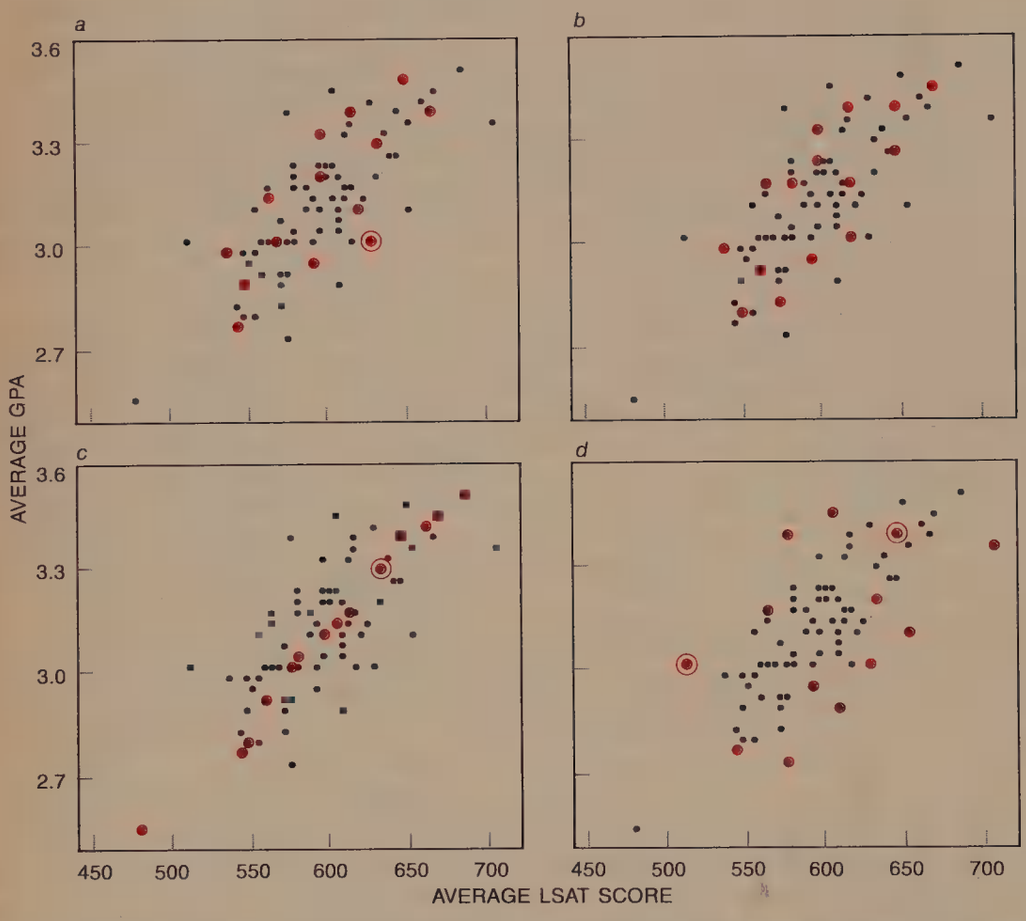
the bootstrap samples, even though the endpoints of the intervals coincide only roughly.

It turns out that the agreement is no accident. Theoretical investigations done by Rudolph J. Beran, Peter J. Bickel and David A. Freedman of the University of California at Berkeley, by Kesar Singh of Rutgers University and by us at Stanford University show that for the correlation coefficient and for a wide variety of other statistics the interval associated with the bootstrap distribution and the interval associated with the distribution of the real samples usually have nearly the same width. (Intervals that include 68 percent of the samples are commonly cited for comparison because for a bell-shaped curve 68 percent of the samples lie within one standard deviation of the peak of the bell.)

At first glance this theoretical result seems paradoxical: it suggests that from the information in each sample one can derive a good approximation to the frequency distribution of the correlation coefficient for all real samples of the same size. It is as if statisticians had discovered the statistical analogue of the hologram, a pattern of light waves that is preserved on a surface. The scene from which the light waves are emitted can be reconstructed in great detail from the whole surface of a hologram, but if pieces of the surface are broken off the entire scene can still be reconstructed from each piece. Not every sample is like a broken piece of a hologram, however; the good properties of the bootstrap are good average properties. Like any other statistical procedure, the bootstrap will give misleading answers for a small percentage of the possible samples.

Suppose the correlation coefficient for the sample of 15 law schools had been nearly 1. That is, suppose all the data points in the sample lay almost perfectly along a straight line. The circumstance is extremely unlikely, given the real data for the 82 law schools, but it could happen. It would then follow that every sample generated by the bootstrap procedure would also lie along the same straight line and so every bootstrap value of r would be nearly equal to 1. The width of the interval associated with 68 percent of the bootstrap samples would therefore be approximately zero. According to the bootstrap procedure, the statistical accuracy of the estimated value of r would be almost perfect, which is incorrect.

The bootstrap does not always guarantee a true picture of the statistical accuracy of a sample estimate. What has been proved is that the bootstrap gives a good picture of the accuracy of the estimate most of the time. There are always a few samples for which the bootstrap does not work, and one cannot know in advance which they are. The limitation is not so much a failure of the bootstrap procedure as it is a restatement of the conditions of uncertainty under which all statistical analyses must proceed.



STATISTICAL ACCURACY of the observed value of r for a random sample can be known precisely only if it is known how r varies for a large number of samples. The 15 law schools for which the value of r has been calculated were selected at random from the total population of 82 American law schools. The data points in each of the four graphs represent the average GPA and average LSAT score for each of the 82 law schools. There are 82^{15} ways to choose samples of 15 law schools from the total population; four such samples have been selected by circling the points in color. (It is possible to select a school more than once in a given sample; such schools have been circled more than once.) The observed values of r for samples a and b are roughly equal to the true correlation coefficient for all 82 schools. The value of r for sample c, however, is much too high and the value of r for sample d is much too low. The true variability in the value of r for samples of 15 law schools can be determined by finding its value for many such samples because data for many more than 15 law schools (in fact, for all 82 of them) are available. Additional data, however, are often impossible to obtain. The bootstrap can estimate the amount of variability that would be shown by all the samples on the basis of one sample.

What are the advantages of applying the bootstrap? In order to appreciate them it is useful to describe how the accuracy of the correlation coefficient (and of most other statistics) was calculated before the computer became widely available. The earlier procedure can be described in terms of the bootstrap, although it goes without saying that before the invention of the computer statisticians did not characterize their methods in such terms. In 1915 the British statistician Sir Ronald Fisher calculated the accuracy of r theoretically. Fisher had to assume that the data for the two variables, average GPA and average LSAT in our example, were drawn at random from a normal probability distribution, represented by a bell-shaped surface. The surface is a two-dimensional analogue of the one-dimensional bell-shaped curve. There is a family of such surfaces whose shape and orientation can be chosen to fit the available set of data. The surface is fitted to the data in the law school sample by placing the top of the bell directly over the point on the graph where both the GPA and the LSAT points have their overall average values. The surface slopes downward to the graph at a rate that depends on how widely the data points are scattered [*see illustration on preceding page*].

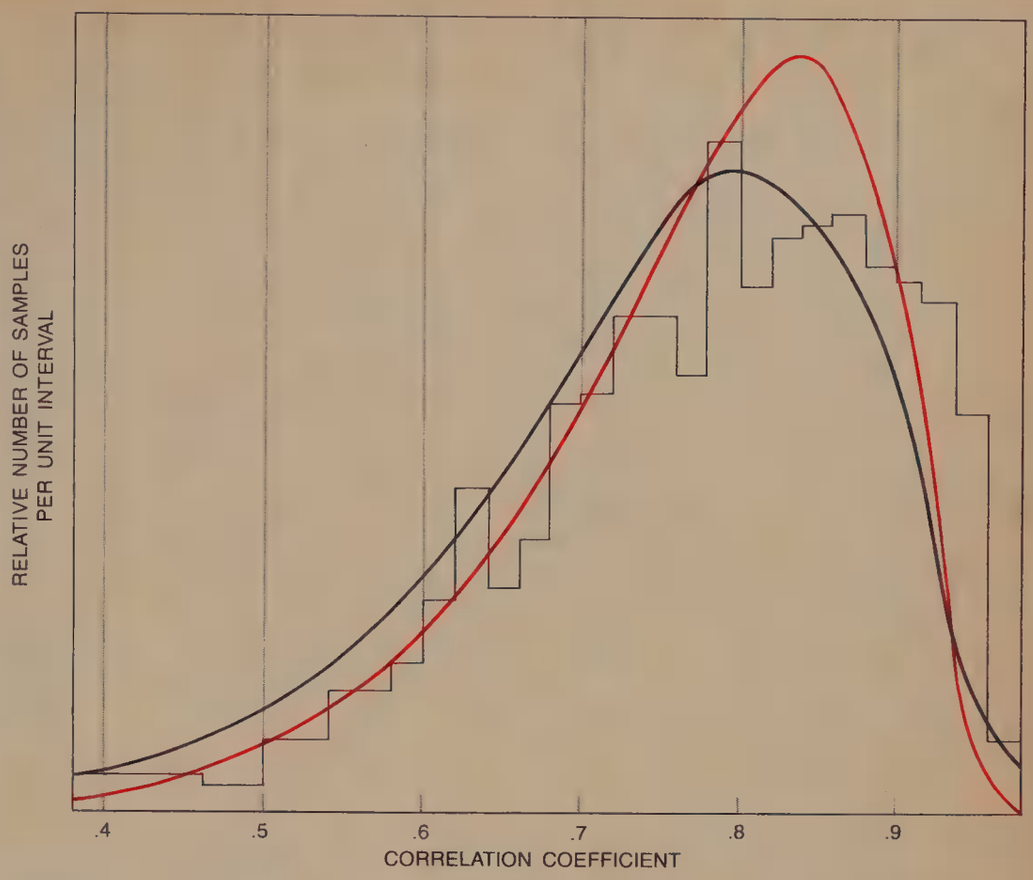
The bell-shaped surface is interpreted as a probability distribution in the same way the graph of values of r for law school samples is a frequency distribution. The probability of selecting a point on the graph of GPA and LSAT scores from within a certain region is equal to the volume that lies under the bell-shaped surface and directly above the region, divided by the entire volume of the space that lies between the surface and the graph. Fisher was then able to

generate a distribution for the values of r by bootstrapping from the bell-shaped probability distribution. In effect, many samples of 15 data points are selected from the graph according to the probability given by their position under the bell-shaped surface. The value of r is calculated for each sample and a frequency distribution for the values of r is plotted. According to Fisher's method, the width of the interval that includes 68 percent of the values of r is .226, in good agreement with the true value of .270 but not as close in this case as the bootstrap estimate of .254.

The bulk of Fisher's calculation can be done analytically because of the assumption that the data in the sample are selected from a normal probability distribution. This assumption is a disadvantage of the method, however, because it might not be true. It is certainly not true in the law school example. Moreover, even if it is true, there is no easy way to check it; in most situations a much larger sample, with perhaps several hundred data points, would be needed to verify the shape of the surface.

The calculations involved in the bootstrap, in which there are no simplifying assumptions about the probability distribution, would have been quite impractical 30 years ago. As we have mentioned, the calculation of a single correlation coefficient takes about five minutes with a desk calculator, and one must carry out from 50 to 1,000 such calculations before a reasonably accurate frequency distribution for the bootstrap samples can be determined.

Today the calculation of a single value of r takes a ten-thousandth of a second with a medium-size computer; at such speed the bootstrap becomes feasible for routine application. If 1,000 bootstrap samples are generated, all the calculations necessary to estimate the width of the interval that includes 68 percent of the samples take less than a second and cost less than a dollar. The cost estimate is based on performing about 100,000 arithmetic operations. More ambitious bootstrap analyses, which give more detailed information about the accuracy of r, require about a million arithmetic operations.



BOOTSTRAP DISTRIBUTION of the correlation coefficient r (*"skyline" in black*) closely approximates the true distribution of r (*smooth curve in black*). The true distribution is actually plotted for a million samples of size 15, chosen at random from the 82^{15} such samples that can be selected from the 82 law schools; differences between the distribution graphed here and the distribution that could in principle be plotted for all 82^{15} samples are not discernible. The shape of the bootstrap distribution also approximates the shape of the distribution that can be estimated according to the probabilities given by a bell-shaped surface (*smooth curve in color*). The agreement suggests the bootstrap can be employed as a measure of the accuracy with which the correlation coefficient of the sample predicts the correlation coefficient of the population. The rather close agreement among the peaks of the distributions is an artifact of the sample.

The bootstrap is not limited to the analysis of the variability of statistics, such as the correlation coefficient, that are mathematically simple. It has been applied to many problems for which the variability of a statistic cannot be assessed analytically. Consider a family of statistics called principal components, which were introduced by Harold Hotelling of Columbia University in 1933. Principal components were devised to solve problems such as the following one, given in a textbook by Kantilal V. Mardia and John T. Kent of the University of Leeds and John M. Bibby of the Open University.

Eighty-eight college students each take two closed-book tests and three open-book tests. Suppose, for the purpose of grading the students, one wants to find the weighted average of the five scores that generates the greatest differences among the students. (In order to make the ratios and not merely the differences of the overall scores as variable as possible, the weights must be scaled so that the sum of their squares is equal to 1.) One set of weights arises if only the final test score is considered; the weights assigned would then be 0, 0, 0, 0 and 1. If all the students had high scores on the final test, however, the summary score generated by this set of weights would not be effective for differentiating the students. Another summary score arises if each test is given equal weight; the weights would then all be equal to $1/\sqrt{5}$, or about .45. The set of weights for the five tests that gives the greatest differences among the students is called the first principal component.

The first principal component is impossible to describe in a mathematically closed form; it must be computed numerically. When the calculation is done for the 88 students, the weights of the first principal component turn out to be roughly equal to one another. The greatest distinctions can therefore be made among the students by finding the average of the five scores.

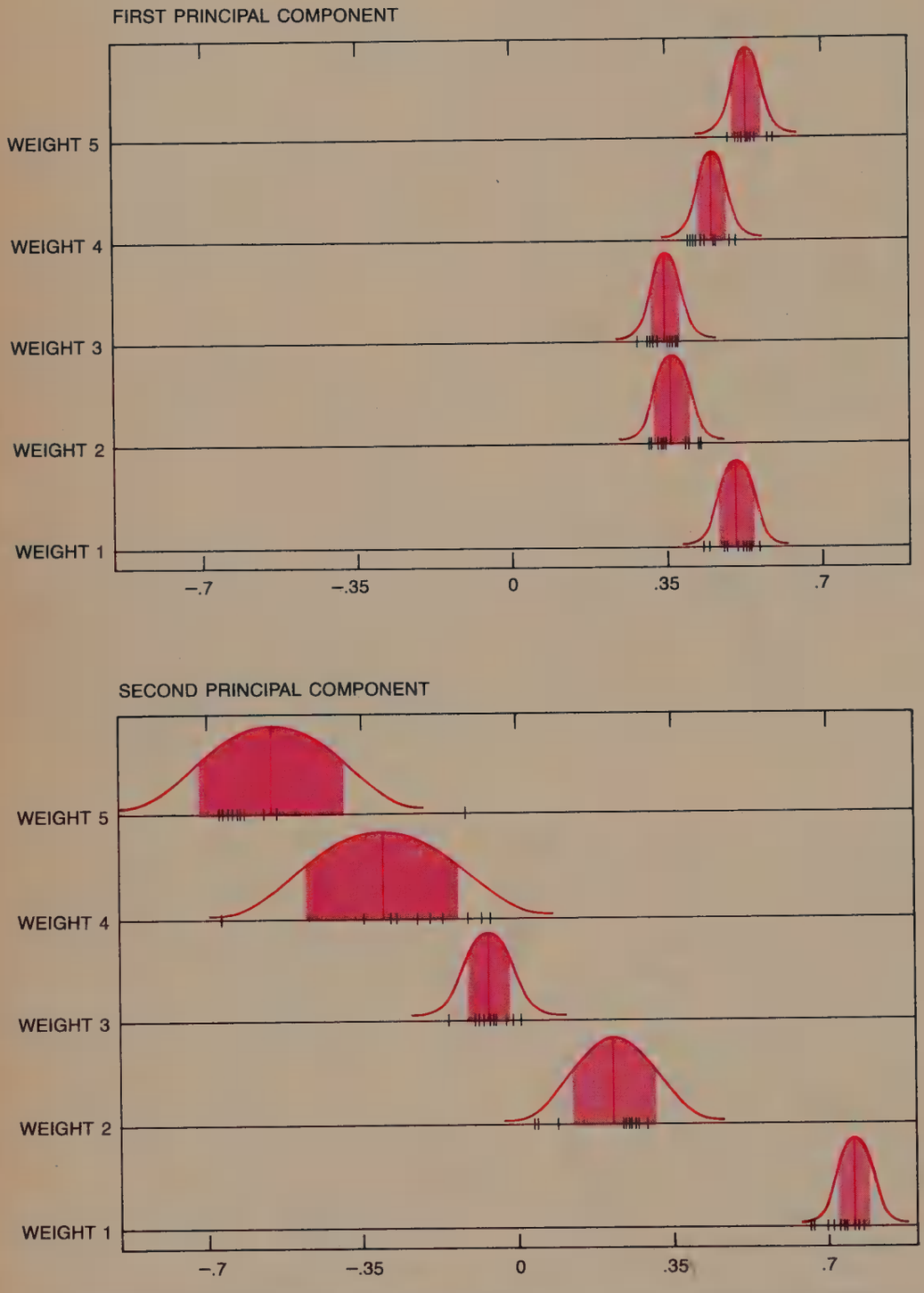
The second principal component is the set of weights, subject to a mathematical constraint of independence, that gives the second-greatest differences among the students. When the second principal component is calculated for the 88 students, the weights turn out to give the difference between an average of the open-book scores and an average of the closed-book scores. The principal components suggest useful and unexpected interpretations for the averages of the student scores. How reliable are the interpretations? If they are to be trusted, one must try to determine how much variation there is in the values of the two principal components for samples of 88 students selected at random.

The problem of quantifying the variability of principal components for samples of a given size has preoccupied many statisticians for the past 50 years. If the appropriate normal distribution is

assumed, partial solutions to questions concerning the frequency distribution for the first principal component can be given; little is known, however, about the second component and higher ones. By applying the bootstrap method a computer can quickly give an estimate of variability for any principal component without assuming that the data have a normal distribution.

In principle the bootstrap analysis is carried out just as it is for the correlation coefficient. Each student's set of five test scores is copied many times (that is, all five scores are written on the same piece of paper) and the copies are thoroughly shuffled. A new sample of size 88 is drawn at random and the principal components are calculated for it. The sampling is repeated many times and a frequency distribution is plotted for each principal component.

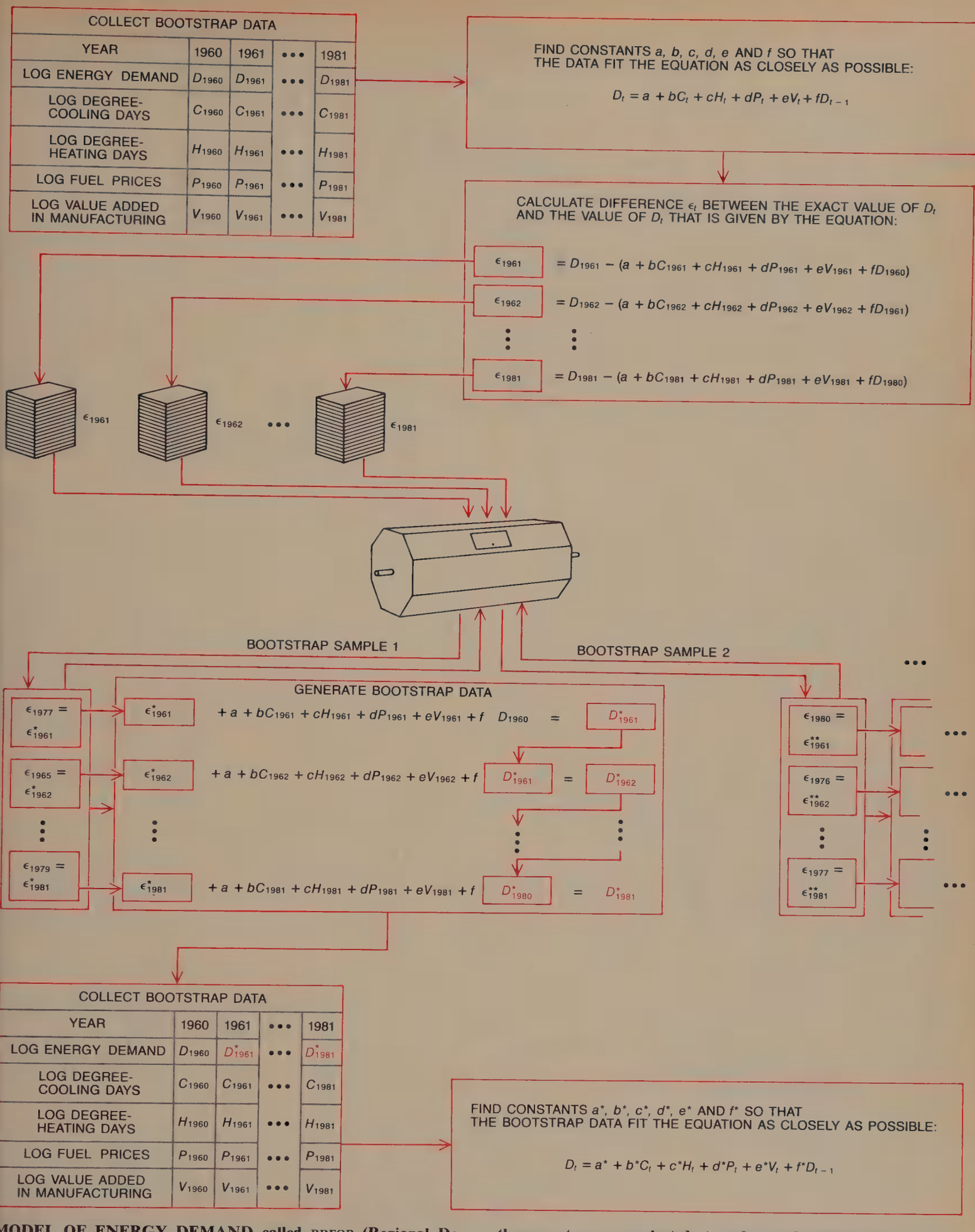
The results suggest that the weights associated with the first principal component are quite stable: they vary only in the second decimal place. The weights associated with the second principal component are less stable, but in a structured way. Remember that the second principal component was interpreted as the difference between an average of the open-book tests and an average of the closed-book tests. The interpretation is confirmed by the bootstrap analysis, but the weights given to the open-book tests are quite variable. The distribution for the principal components generated by the bootstrap is a good estimate of the true distribution of the principal components for samples of size 88. It takes only about two seconds for a large computer to do 100 bootstrap replications.



PRINCIPAL COMPONENTS are statistical estimators that have been widely applied for calculating summary scores on standardized tests. Suppose 88 students take five tests each and suppose, in order to assign a summary score, one wants to find the weighted average of the five test scores that generates the greatest differences among the students. The first principal component is the set of weights that solves the problem. The second principal component is the set of weights, subject to a mathematical constraint of independence, that generates the second most variable combination of the test scores. To assess the variability of the two principal components for many additional samples of 88 students the bootstrap was applied to the single sample. Each student's score for the five tests was written on a piece of paper, and each set of five scores was copied many times. All the copies were then shuffled and bootstrap samples of size 88 were selected at random. The principal components were calculated for each bootstrap sample. The variation in the weights for the first 10 bootstrap samples is shown by the black ticks on each graph; the red vertical lines indicate the observed values of the weights. The width of the central strip under the small bell-shaped curves indicates the variability of the weights. The fourth and fifth weights of the second principal component are particularly unstable.

Not every statistical estimator is a number. Nine weather stations in the eastern and midwestern U.S. recorded the *p*H level, or acidity, of every rainfall from September, 1978, through August, 1980. (A *p*H value of less than 7 is acidic, and the lower the *p*H value, the greater the acidity.) During the two years 2,000 *p*H values were measured. To represent the data Barry P. Eynon and Paul Switzer of Stanford prepared a *p*H-contour map of the region; the *p*H values are constant along a contour line. Such a map can be generated from the data by a well-defined mathematical procedure called Kriging, after the South African mining engineer H. G. Krige. Although the contour map is strictly determined by the data, it represents an extrapolation from the data collected at nine stations to many points in space and time (in fact, to an infinite number) that are not included in the original sample. One can therefore ask how variable the contours on the map would be if random variations yielded different samples of 2,000 *p*H values.

In this example neither the true contour map nor the true variability of all contour maps generated by samples of 2,000 *p*H values can be known. Both estimates must be made from the original data alone if they are to be made at all. By bootstrapping the original sample of 2,000 *p*H values in a way that preserves the geographic relations among the weather stations, Eynon and Switzer generated the maps shown in the illustration on page 117. There is no generally accepted measure of the variability of contour lines on a map, analogous to the width of an interval of a frequency distribution. Intuitively, however, the variability is readily perceived. It shows that the original contour map must be inter-



MODEL OF ENERGY DEMAND called RDFOR (Regional Demand Forecasting Model) was employed by the U.S. Department of Energy to analyze and forecast energy demand in 10 regions of the U.S. For each region the data are fitted as closely as possible to a mathematical model called a regression equation. The demand for energy in any given year is assumed to depend on the demand for energy the year before as well as on several other measures. Each error term ϵ_t is the difference between the predicted value of the energy demand in a given year and the observed value. Bootstrap samples of the error terms are selected at random and artificial data for energy demand are generated by the method shown in the diagram. The bootstrapped data are then fitted to a new regression equation, and the variability of the regression equations generated by the bootstrap gives an estimate of the expected accuracy of the model in predicting energy demand. A bootstrap analysis done by David A. Freedman of the University of California at Berkeley and Stephen C. Peters of Stanford has shown that the variability of the regression equations is from two to three times greater than was previously thought.

preted cautiously. Corridors of relatively low or relatively high acidity on the original map can shrink to become islands on a bootstrap map, depending on the effects of random "noise."

Statistical estimation is often carried out by making the available data conform as closely as possible to some predetermined form, or model. The simplest models are the line, the plane and the higher-dimensional analogues of the plane. Consider the graph of the 15 data points that represent the 15 law schools. Intuitively there are many lines that could be drawn to represent the trend of the data points, and so it is reasonable to agree in advance on a precise method for fitting the points to a line. Probably the most widely used estimator in statistics is a method for fitting points to a line called the least-squares method. The method was invented by Gauss and by Joseph Louis Lagrange in the early 19th century in order to make astronomical predictions.

The least-squares line is the line that minimizes the sum of the squares of the vertical distances between the data points and the line. A straightforward calculation gives the equation of the least-squares line from the data points. If the bootstrap is applied to the data, fake data sets can be generated, and the least-squares method can be applied to each fake set of data points in order to fit them to a new line. The fluctuation of the lines generated by the bootstrap shows the variability of the least-squares method as a statistical estimator for this set of data points.

	PATIENT NUMBER						
	149	150	151	152	153	154	155
AGE	20	36	46	44	61	53	43
SEX	M	M	M	M	M	F	M
PRESENCE OF STEROID?	YES	NO	NO	NO	YES	YES	NO
ANTIVIRAL ADMINISTERED?	NO	NO	NO	NO	NO	NO	NO
FATIGUE?	NO	NO	YES	YES	YES	YES	YES
MALAISE?	NO	NO	YES	NO	YES	NO	NO
ANOREXIA?	NO	NO	YES	NO	NO	NO	NO
LARGE LIVER?	NO	NO	NO	NO	YES	NO	NO
FIRM LIVER?	*	NO	NO	YES	YES	NO	NO
PALPABLE SPLEEN?	NO	NO	NO	NO	NO	YES	YES
PRESENCE OF SPIDERS?	NO	NO	YES	NO	YES	YES	YES
PRESENCE OF ASCITES?	NO	NO	YES	NO	NO	NO	YES
PRESENCE OF VARICES?	NO	NO	YES	NO	NO	YES	NO
CONCENTRATION OF BILIRUBIN	.9	.6	7.6	.9	.8	1.5	1.2
CONCENTRATION OF ALKALINE PHOSPHATASE	89	120	*	126	95	84	100
CONCENTRATION OF SERUM GLUTAMIC-OXALOACETIC TRANSAMINASE (SGOT)	152	30	242	142	20	19	19
CONCENTRATION OF ALBUMIN	4.0	4.0	3.3	4.3	4.1	4.1	3.1
CONCENTRATION OF PROTEIN	*	*	50	*	*	48	42
PHYSICIAN'S PROGNOSIS	LIVE	LIVE	*	LIVE	LIVE	*	LIVE
OUTCOME	LIVE	LIVE	DIE	LIVE	LIVE	LIVE	DIE

MEDICAL DATA for seven out of 155 people with acute or chronic hepatitis give the values of 19 variables for each person that, taken together, could predict whether a patient will die or recover from the disease. (An asterisk indicates that information is missing.) It is common practice in statistics to inspect such data before a formal model is constructed; the aim of the inspection is to rule out all but four or five of the most important variables. Peter B. Gregory of the Stanford University School of Medicine eliminated all the variables except the patient's malaise, the presence of ascites (fluid in the abdominal cavity), the concentration of bilirubin and the physician's prognosis. Gregory developed a model based on the four variables that correctly predicted whether or not the patient would die from the disease in 84 percent of the cases.

The least-squares method and its generalizations are particularly useful for complex problems in which an investigator must bring large amounts of diverse information to bear on a single question. The U.S. Department of Energy, for example, has developed a model called the Regional Demand Forecasting Model (RDFOR), which attempts to forecast the demand for energy in 10 regions of the U.S. In the model it is assumed that the energy demand for each region in a given year depends in a simple way on five variables: the amount of variation above 75 degrees F. in summer, the variation below 65 degrees in winter, the price of fuel, the value added in manufacturing (a measure of the economic conditions in the region) and the energy demand during the previous year.

The five variables can be thought of as if they were plotted on a five-dimensional graph, which is exactly analogous to a two-dimensional graph; every point on a five-dimensional graph corresponds to a possible combination of the five variables. The energy demand in a given year associated with a known combination of variables can then be represented by the height of a point in a six-dimensional space above the corresponding point on the five-dimensional graph. The representation of the data in a six-dimensional space is analogous to representing the dependence of some quantity on two other variables as the height of a point in three-dimensional space above a two-dimensional graph. Thus the energy data determine a set of points at various heights in the six-dimensional space.

The least-squares method specifies a way of drawing a five-dimensional analogue of a plane (called a hyperplane) as close as possible to all the points. Because of the dependence of the energy demand on the demand in previous years, the variables must be fitted to the hyperplane by a generalized version of the least-squares method. The generalized method calls for minimizing a weighted sum of errors after the weights have been estimated from the data. In recent years an elaborate method of estimating the accuracy of the procedure and the accuracy of its forecasts has been developed.

Freedman and Stephen C. Peters of Stanford examined the conventional estimates of the accuracy of the procedure by applying the bootstrap. In their approach it is assumed that the data lie close to an appropriate hyperplane, but it is not assumed that the errors between the data points and points that lie on the hyperplane are independent of one another. Instead the relation of the errors from point to point is allowed to have a complicated structure. The bootstrapping of the data was done in a way that

© 1982 Toyota Motor Sales, U.S.A., Inc.

Grand Touring. What does it mean?

In the new 1983 Toyota Cressida, Grand Touring means Grand Performance. With an amazing new 2.8 liter Twin Cam engine that does to smoothness and quietness of ride, what Michelangelo did to a ceiling. Add that to Cressida's new independent rear suspension, with coil springs and stabilizer bar, and its new electronically controlled 4-speed automatic overdrive transmission and you start to see how grand a driving experience can be.

Cressida's Grand Touring also means Grand Luxury. From the look of luxury outside — dashing, elegant, refined — to the feel of luxury inside — power windows and door locks. Automatic temperature control air conditioning. Variable assist power steering. Cruise control. AM/FM/MPX stereo receiver with four speakers. And a driver's seat that adjusts to your body in four distinct ways.

What else does Grand Touring mean?

In the new Cressida, it represents a feeling you get, while touring town or country. A feeling based on uncompromised performance. And spirited drive. It's the image of the car you've chosen. And of yourself.

Dashing.

BUCKLE UP—IT'S A GOOD FEELING!

OH WHAT A FEELING!

TOYOTA

CRESSIDA MOVES PROUDLY INTO THE GRAND TOURING CLASS.

DASH!

preserves the evolution of the energy demand from year to year. The variability of the hyperplanes generated by the bootstrap showed that the standard error previously assumed for the energy model is usually too small by a factor of two or three. The predictions of energy demand made by this aspect of the RDFOR model are therefore much less reliable than was once thought.

The examples we have presented so far have involved clearly defined statistical properties of samples. In practice the data can be inspected, sorted, graphed and preanalyzed in several ways before they are formally analyzed. Estimates of variability that do not take such informal practices into account may not give an accurate picture of statistical variability.

Consider a group of 155 people with acute and chronic hepatitis, initially studied by Peter B. Gregory of the Stanford University School of Medicine. Of the 155 patients 33 died and 122 survived, and for each patient there were data for 19 variables, such as age, sex and the results of standard biochemical measurements. Gregory's aim was to discover whether the data could be combined in a model that could predict a patient's chance of survival.

The analysis of the data was done in several steps. First, all but four of the most important variables were eliminated, because statistical experience suggests it is unwise to fit a model that depends on 19 variables with only 155 data points available. The elimination of the variables was done in two stages: each variable was inspected separately, whereupon six variables that appeared unrelated to the patients' survival were eliminated. A standard statistical procedure was then carried out on the remaining 13 variables, which further reduced the number to four. The variables that remained were the patient's malaise, ascites (the presence of a fluid in the abdomen), the concentration of bilirubin in the liver and the physician's prognosis for the patient. The variables were then fitted to a curve that predicts how the proportion of surviving patients depends on the values of the variables.

Such analysis is typical of scientific practice. In order to estimate its overall variability Gail Gong of Carnegie-Mellon University carried out the entire procedure from preliminary screening through the final curve fitting on bootstrapped samples of the original 155 data points. Her results were surprising and informative. The set of "important variables" generated during the initial stages of the analysis was itself quite erratic. For some bootstrap samples only the prognosis of the physician was found to be important, whereas for others such variables as sex, age, level of fatigue, concentration of albumin and concentration of protein were retained. No single variable emerged as significant in as many as 60 percent of the bootstrap samples.

Although the fitted curve is intended to predict whether or not a patient will survive, it misclassifies 16 percent of the original 155 patients. The estimate of 16 percent, however, is too small because the data on which it is based were also employed to generate the curve. The analysis generated by the bootstrap suggests a better estimate for the probability that the fitted curve will misclassify a given patient is .20.

The prospect of bootstrapping the entire process of data analysis offers hope that an extremely difficult problem will begin to yield, namely the connection between the mathematical theory that underlies statistics and actual statistical practice. The effects of preliminary "data snooping" on the final results are usually ignored, for no better reason than that it is impossible to analyze them mathematically. It now appears that the bootstrap, applied with the aid of the computer, can begin to estimate such effects.

The bootstrap is by no means the only statistical method that relies on the power of the computer. Several other methods such as the jackknife, cross-validation and balanced repeated replications are similar in spirit to the bootstrap but quite different in detail. Each of these procedures generates fake data sets from the original data and assesses the actual variability of a statistic from its variability over all the sets of fake data. The methods differ from the bootstrap and from one another in the way the fake data sets are generated.

The first such method was the jackknife, invented in 1949 by Maurice Quenouille and developed in the 1950's by John W. Tukey of Princeton University and the Bell Laboratories; it has been extensively investigated by Colin L. Mallows of Bell Laboratories, Louis Jaeckel of Berkeley, David V. Hinkley of the University of Texas at Austin, Rupert G. Miller of Stanford, William R. Schucany of Southern Methodist University and many others. The name jackknife was coined by Tukey to suggest that the method is an all-purpose statistical tool.

The jackknife proceeds by removing one observation at a time from the original data and recalculating the statistic of interest for each of the resulting truncated data sets. The variability of the statistic across all the truncated data sets can then be described. For the data from the 15 law schools the jackknife assesses the statistical accuracy of the value of r by making 15 recalculations of r, one for every possible subsample of size 14. The jackknife calls for fewer calculations than the bootstrap but it also seems less flexible and at times less dependable.

Cross-validation is an elaboration of a simple idea. The data are split in half and the second half is set aside; curves are fitted to the first half and then tested one by one for the best fit to the second half. The final testing is the cross-validation; it gives a reliable indication of how well the fitted curve would predict the values of new data. There is nothing special about half splits; for example, the data can be split in the ratio 90 to 10 as

BOOTSTRAP SAMPLE NUMBER	VARIABLES SELECTED
491	ALBUMIN, PROGNOSIS, SEX
492	ASCITES, BILIRUBIN, PROGNOSIS
493	BILIRUBIN, ASCITES
494	BILIRUBIN, PROGNOSIS, MALAISE
495	ASCITES
496	BILIRUBIN
497	ASCITES, VARICES
498	SPIDERS, PROGNOSIS, ALBUMIN
499	AGE, PROGNOSIS, BILIRUBIN, MALAISE, PROTEIN, SPIDERS
500	ASCITES, PROGNOSIS, BILIRUBIN, PROTEIN

VARIABLES DESIGNATED IMPORTANT by informal analysis prior to the construction of a formal statistical model can show wide variation. In a bootstrap study that simulated both the formal and the informal aspects of the statistical analysis, Gail Gong of Carnegie-Mellon University programmed a computer to copy the set of 19 variables associated with each patient many times. The sets of data were thoroughly shuffled and bootstrap samples of 155 sets of data were drawn at random from the collection. Formal and informal techniques of data analysis were then applied to each bootstrap sample, just as they had been for the original sample. The variables chosen as important are shown for 10 of the 500 bootstrap samples generated. Of the four variables originally chosen not one was selected in as many as 60 percent of the samples. Hence the variables identified in the original analysis cannot be taken very seriously.

Exxon has the information processor...

that really can be the start of something big.

At Exxon Office Systems, we're bringing the high-tech office down to earth, by designing office automation to grow the way you grow.

You can start an office automation system simply with just one of our remarkable EXXON 500 Series Information Processors.

Then as you grow, we can grow with you, right into our shared resource office automation system ... The EXXON 8400 Series System.

As a fully functioned system its capability is awesome. Operators can create, edit, reformat, file, share and retrieve documents, all with the touch of a key.

There's a dictionary, an electronic mailbox, a program for keeping calendars and scheduling meetings, a tickler file and more.

Start automating your office now with one of our EXXON 500 Series Information Processors. And step into the future without the shock of unnecessary costs from expanding inefficiently.

For more information on the EXXON 500 or our new office automation system, just send in the coupon below. Or call **800-327-6666**.

EXXON OFFICE SYSTEMS

The future_without the shock.

Exxon Office Systems SAM 05 83
P.O. Box 10184, Stamford, CT 06904

I'd like to know more about

☐ the EXXON 8400 Series System
☐ the EXXON 500 Series Information Processor

Please have your representative call.

Name _____ Title _____

Company _____

Address _____

City _____ State _____ Zip _____

Telephone _____

800-327-6666.

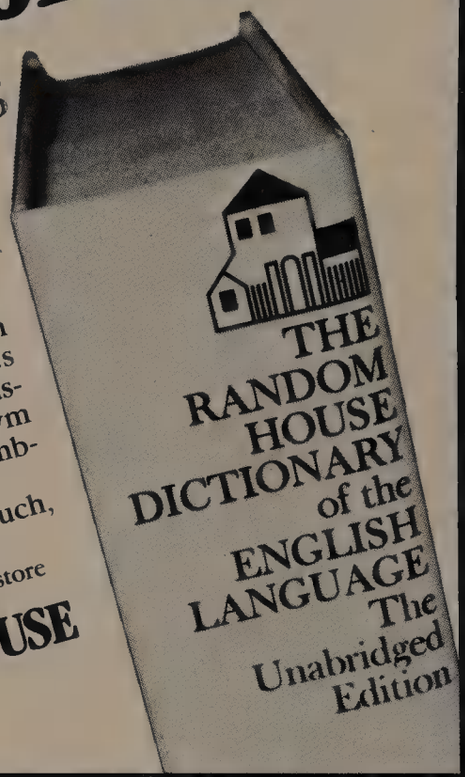
IN CONNECTICUT, 800-941-3535.

THE Answer Book
The Unabridged!

- THE MOST AUTHORITATIVE DICTIONARY OF ITS KIND
- Comprehensive—more than 260,000 entries; 2,091 pages; large format 9" x12" page size; 9 lbs 14 oz.
- Up-To-Date—with new words and terms
- Easy to Use—more than 50,000 example phrases and sentences; 2,000 illustrations; 10,000 synonym lists and studies; thumb-indexed
- Full-color ATLAS; much, much more

$49.95, now at your bookstore

RANDOM HOUSE



Why has higher intelligence been sparked in only one species?

Promethean Fire Reflections on the Origin of Mind

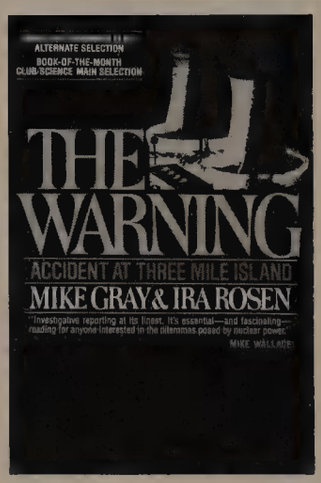
Charles J. Lumsden
Edward O. Wilson
Illustrations by Whitney Powell

"*Promethean Fire* ponders the virtually imponderable, the origin of our own minds, and comes out on top. Every page is an adventure, the concepts come at you like laser beams. A thoroughly provoking book."
—Roger A. Caras

$17.50 *Illustrated*

At bookstores or order direct.
Harvard University Press
Cambridge, MA 02138

"If you read no other book in the next five years, read THE WARNING!"
L.A. Times Book Review



THE WARNING: ACCIDENT AT THREE MILE ISLAND was written by Mike Gray, author of the film THE CHINA SYNDROME, and Ira Rosen, a producer of 60 MINUTES.

Now available in paperback

Contemporary Books, Inc.

well. Moreover, there is no reason to carry out the cross-validation only once. The data can be randomly split many times in many ways.

Cross-validation has been widely applied to situations in which a curve-fitting procedure is well defined except in some crucial respect. For example, one might be willing to fit a polynomial to the data by the least-squares method, but the degree, or highest power, of the polynomial to be fitted might still be in doubt. (The higher the degree of the polynomial, the less smooth the fitted curve.) Given that half of the data have been fitted by polynomials of various degree, cross-validation can choose the degree of the polynomial that best fits the second half of the data. Seymour Geisser of the University of Minnesota, Mervyn Stone of the University of London and Grace G. Wahba of the University of Wisconsin at Madison have been pioneers in this development.

Instead of splitting the data in half at random a more systematic system of splits can be employed. The splits can be chosen in such a way that the results are optimal for certain simple situations that allow full theoretical analysis. The balanced repeated-replication method, developed by Philip J. McCarthy of Cornell University, makes splits in the data systematically in order to assess the variability of surveys and census samples. Random subsampling, a related method developed by John A. Hartigan of Yale University, is designed to yield dependable confidence intervals in certain situations.

There are close theoretical connections among the methods. One line of thinking develops them all, as well as several others, from the bootstrap. Hence one must ask what assurance can be given that the bootstrap will work most of the time, and how much it can be generalized. To the first question the answer is simple. The bootstrap has been tried on a large number of problems such as the law school problem for which the correct answer is known. The estimate it gives is a good one for such problems, and it can be mathematically proved to work for similar problems.

We have suggested the answer to the second question through the diversity of complex problems to which the bootstrap has already been applied. What is needed for many of them, however, is independent theoretical justification that the bootstrap estimate of accuracy remains as valid as it is for simpler problems. Current theoretical work seeks to provide such justification and to give more precise statements of accuracy based on the bootstrap. Fisher was able to provide a statistical theory that took full advantage of the computational facilities of the 1920's. The goal now is to do the same for the 1980's.

6

Sixth in a series of how Delco Electronics and Bose technology contribute to your enjoyment of driving.

Deceptive...isn't it.

The control panel of this Electronically Tuned Receiver (ETR) is simple—and deceptive. Simple so that the receiver is easy to operate. Deceptive because a very sophisticated technology lies behind it. A technology that produces high fidelity reception from the Delco-GM/Bose Music System under conditions that are even difficult for ordinary radio reception.

The key to this technology is Delco Electronic's own custom integrated circuits. These circuits respond automatically to changing reception conditions and program requirements. So you can enjoy music and driving more.

When you visit your GM dealer* you will understand why Len Feldman wrote in Popular Science: "It's as good as or better than the best home systems I've heard."

*Available as a factory-installed option on Cadillac Seville and Eldorado, Buick Riviera, Oldsmobile Toronado and Corvette by Chevrolet.

Delco GM BOSE

Sound so real it will change how you feel about driving.

The Social Influence of the Motte-and-Bailey Castle

The motte was a medieval mound fort; the bailey was an associated enclosure. Their appearance in the second half of the 10th century diminished central authority in Europe and gave rise to chivalry

by Michel Bur

The first project of military engineering ordered by William the Conqueror after his landing in England was the hasty construction of a motte, a fortification of a kind unknown in the British Isles. Designed to provide the Normans at Hastings with a strongpoint, the motte was primarily a mound of earth surrounded by a ditch. Embedded in the mound were heavy timbers that supported a tower put together out of prefabricated members. In the years that followed the Norman victory the occupiers built similar mottes throughout England, thus securing a firm hold over the vanquished.

The Normans brought their knowledge of mottes with them from France, where the building of such simple fortifications had enabled nobles, great and small, to defy the central authority of Charlemagne's successors and to establish their own domains. It is not too much to say that this novel defensive weapon revolutionized medieval Europe. Politically it was the instrument of the feudalism that replaced rule from a single center. Socially its effects were perhaps even more important: the motte served as a kind of schoolroom where the medieval seigneurs and their retinue learned what was needed to create the form of civilized behavior known as chivalry.

Until a few years ago little was known about the mottes of France. The only archaeology concerned with the Middle Ages was the study of major medieval buildings, a subdiscipline of history concerned with describing and analyzing the technical aspects of masonry construction. The examination of other remains of medieval life was left to geographers; archaeologists did not bother to probe them. Recently, however, scholars at universities and other institutions have begun investigating the medieval countryside, and their work has begun to transform long-established patterns of historical thought.

Among the areas where progress in such research is being made is the study of fortified rural habitations. The motte has proved to be the most significant of these primitive seigneurial strongholds of the late 10th century through the 12th. Thanks to archaeological studies outside France, particularly those in Britain, where the details of the Norman Conquest continue to be investigated, French defensive earthworks of the period are well known.

A typical motte, appearing in profile either as a truncated cone or as part of a sphere, could be up to 100 meters in diameter at its base and as much as 20 meters high. Adjacent to most mottes was a large earth-walled enclosure, known as a bailey, surrounded by a ditch and topped with a wood palisade. Hence this type of dual defensive earthwork is called a motte-and-bailey castle. Another type of earthwork consists of a miniature bailey without a motte: a circular or oval bank 30 to 100 meters in diameter, surrounded by a ditch and topped with a palisade. Some of these enclosures served only as pens for livestock. Others, particularly those flanked by separate stock pens, were actually earthen castles of a lesser type.

Since 1961 a systematic inventory of these primitive seigneurial habitations has been under way in France, beginning in the Auvergne. Normandy was added to the survey in 1968 and Champagne in 1972. Since then the work has been extended both eastward and westward to include French Flanders, Dauphiné, Gascony and Saintonge. Its objective is to find, describe, measure and map (on a scale of 1 : 1,000) every vestige of these medieval earthworks that remains before modern agricultural machinery has erased them. The field work is being coordinated with archival research aimed at determining the date of construction and of abandonment or razing, the name given to each motte (such as "stronghold" or "castle") and the role in society of its inhabitants, such as "count" (*comes*), "knight" (*miles*) or "seigneur" (*dominus*). The standardization of such inquiry should result in a uniformity of collected data and so provide, if need be with the help of computer programs, a solid quantitative base for historical speculation.

A mere sampling of sites would not be acceptable. The survey must, like oil poured on water, spread outward from one community to the next until the census is complete. In brief, the study, both methodical and exhaustive, is simultaneously examining in depth the medieval archives and what remains of the medieval landscape.

The motte-and-bailey castle was known as a *château,* a word that like "castle" is from the Latin *castellum.* It was an establishment that served as a permanent abode for a seigneur and his men-at-arms. The *château* differed in size and function from similar but much larger defensive enclosures that had in the past served as places of refuge for the general population. The distinction is of particular importance because the motte, from the military point of view, was the most significant technological innovation of the latter part of the 10th century.

Mottes appeared first in low-lying marshy areas in the region between the Loire and the Rhine and next on the steep hills of the same region. From this "homeland" they spread across Europe from the Atlantic to the Vistula and from Scandinavia to the Mediterranean. With their wood tower—in some instances simply a watchtower, in others a larger residential structure—these earth mounds were, with one significant difference, equivalent to the nearly contemporaneous great stone donjon keeps of the Loire and Normandy. The difference was that in western Europe the medieval mottes were much more numer-

ous than the stone castles were then or at any time later.

The time when the mottes quickly spread across Europe was an eventful one. Even before A.D. 1000 parts of western Europe (at first France and Italy and later Germany) had undergone a profound political and social transformation. The structure of Charlemagne's empire was in ruins and smaller political entities were forming everywhere. In place of the Carolingian system of state control from the center, men grouped themselves around feudal lords. The locally powerful who were sought out as protectors by lesser men now intended to exercise for their own benefit the authority they had previously drawn from the king. It was a revolution of the landed aristocracy, the very group from whose ranks the sovereign had traditionally recruited his administrators and officers. In France the landed aristocracy eventually dethroned the descendants of Charlemagne and in A.D. 987 gave the crown to one of their own, Hugh Capet. In the years that followed the interests of the landed aristocracy stifled those principles of peace and common weal that had formerly had the support of the central authority.

Such an upheaval could not take place without fierce strife: the great nobles against the monarch, the same nobles among themselves and so on down to second- and third-tier seigneurs against one another. The 10th century was a time of particularly violent competition for land and power. If a seigneur was to dominate, he needed abundant resources: the means to maintain troops and reward the fidelity of the rank and file. In the long run, however, even the most determined adventurer was always at risk of chance defeat, perhaps only a temporary setback, perhaps death in a skirmish. No territorial takeover could have been lasting without the motte, a weapon that enabled those who controlled it to defy their adversaries and to transform that defiance into irreversible victory.

Just as the waterwheel mill and the draft-animal harness revolutionized the economics of agriculture, so was the motte an instrument of political and social revolution. An unskilled peasant labor force could easily erect the earthworks in a short time. Moreover, except for its wood superstructure the motte was virtually indestructible. When in the 11th century provincial authority (itself a product of the breakdown of central authority) was challenged in such strongholds as Normandy and Champagne, what followed? Mottes sprang up all over the two provinces; without that weapon those up in arms against

LARGE MOTTE in the Ardennes of northeastern France is seen to the left of the church in this aerial photograph of Cornay, a village near the canal town of Vouziers. Although eroded for more than seven centuries, this earthwork is still 18 meters high, 60 meters in diameter at the base and 35 meters in diameter at the top. The summit, which covers an area of 110 square meters, now serves as a cemetery.

the provincial authorities could neither have initially held their ground nor have eventually sunk their roots.

Motte tactics were not unlike those of a huge but simplified game of chess. The primary objective was to advance one's own pieces without having them captured. One pushed forward and built a motte, strongly barricading one's forces within it to resist the siege that one's rival, hoping to regain lost lands, was sure to begin immediately. In general such sieges were failures, and after the seigneur and his garrison had survived a few such efforts the motte, by now upgraded to the status of a *château,* was guaranteed to endure at least until treason opened the gates or the seigneur himself was captured.

Contemporaneous accounts show that the early seigneurs lived in constant fear of treason, never being entirely sure of their followers' fidelity. They had to avoid capture at all costs, a consideration that explains why, when things went adversely in formal open battles, the seigneur in command of the losing side was the first to leave the field. The reason was that once the loser was in the hands of the winner he was exposed to intolerable pressure, as were his kin and others left behind to guard his domain. Many a captive seigneur was brought to the door of his *château,* to be hanged there if his wife did not open the gates. Many a noble who held as hostages the children of a traitorous seigneur threatened to hang them or blind them if their father did not give himself up. Less dramatic than such episodes but fully as effective in putting pressure on a captive was one involving the comte de Blois in 1044. After his capture by the comte d'Anjou a mere three days of imprisonment convinced him to yield the city of Tours and its surrounding countryside in exchange for his freedom.

Two schools of medieval historians disagree on how the mottes came to proliferate so rapidly. One school takes the view that most such *châteaux* were legally built by those endowed with authority. The other, echoing a phrase coined by Norman chroniclers, sees almost every motte as a *château adultérin,* that is, an unlawful construction. In effect anyone who declares some action to be a revolution necessarily defines as unlawful the means by which the revolution triumphs. Actually whether most of the *châteaux* were built by properly authorized persons or more informally by

MOTTE AT HASTINGS, the site of William the Conqueror's battle with Harold, appears in a detail from the Bayeux tapestry. The corner posts of the tower have been put up and are being covered with fresh earth by two men with pickaxes and three with shovels. For major operations such as this one and perhaps even for lesser operations the motte builders carried with them timbers cut to size.

adventurous landed proprietors matters little since what counts is not so much the source of the power as the purpose to which the power was put. This argument leads to a complementary one. If the act of revolution renders unlawful all other actions against the challenged authority (in this instance a violation of the monopoly on fortifications held by the crown or by lesser nobility), the revolt, if it is successful, soon creates its own legality. One is reminded of a 16th-century epigram of John Harington: "Treason doth never prosper: what's the reason? / For if it prosper, none dare call it treason."

In a military aristocracy social relations were based on the vassal's homage and oath of allegiance to his superior. In other words, the weak became dependent on the strong in return for a reward in the form of land: the fief. This principle, occupying the middle ground between strength and weakness, gradually became the criterion of a new kind of legality that applied equally to kings and to great nobles: to be accepted as a legal occupant the master of a *château* had to hold the fortification in fief from a sovereign or a nobleman. In the give-and-take that naturally put the holders of existing fortifications in opposition to those who sought to build new ones, an equilibrium began to develop. This situation owed nothing to a policy of direct confrontation, an always violent and often sterile course of action. It came about because of recourse to the legal formula of feudalism. For the more politically skilled nobles the building of a stable establishment called not for destroying (or trying to destroy) the *châteaux adultérins* but for bringing them into vassalage. Thus the mottes on the fringe of any particular jurisdiction were progressively transformed into "feudal" mottes, with their existence no longer contested.

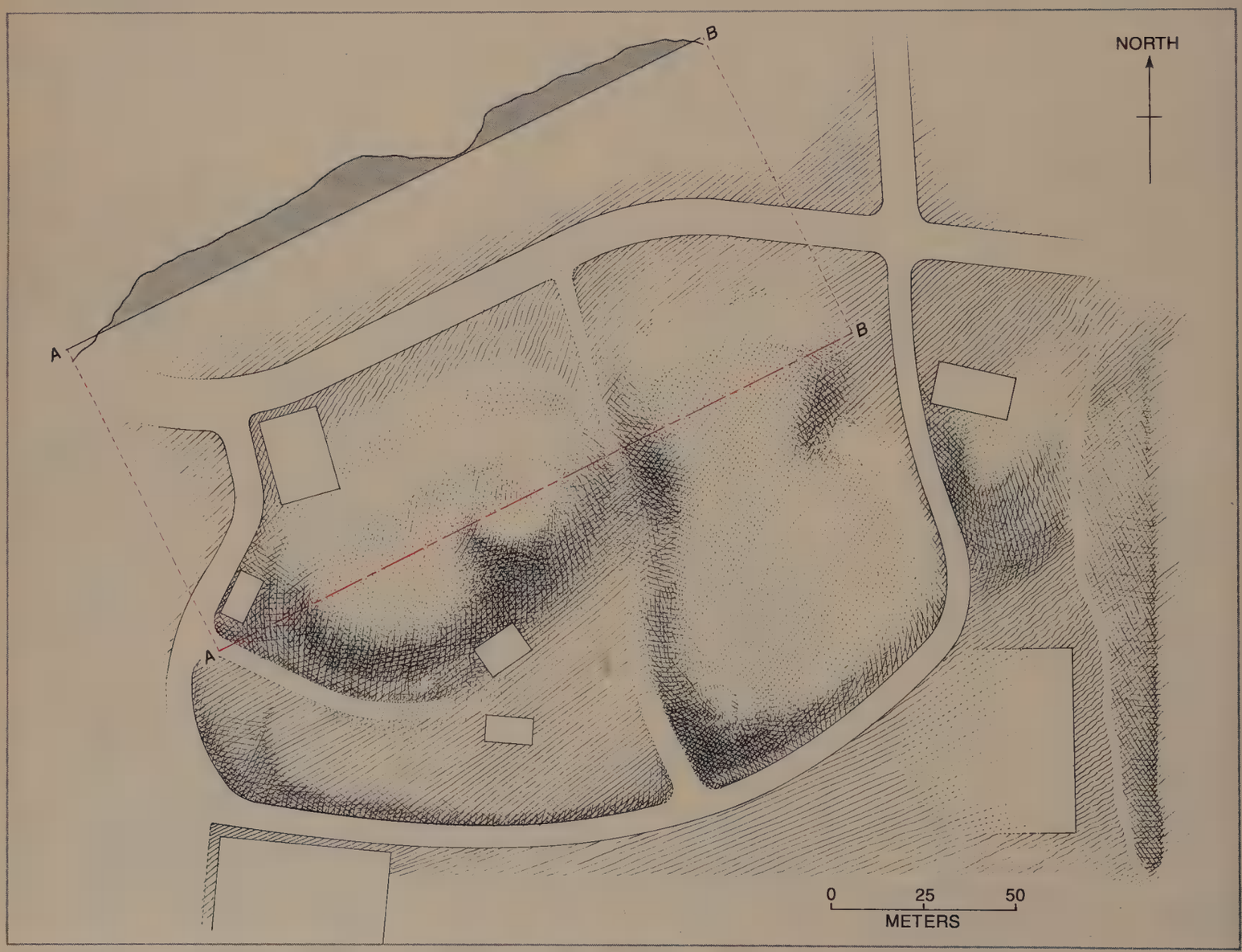
The network of the first French mottes had no connection with existing population centers; the formidable mass of the palisaded fortress was raised in total disregard of the local population. Keeping close watch on their rivals, each seigneur's forces camped in the flat countryside like an army of occupation, their ties with the countryfolk limited to levies and labor drafts. Above all, the *château* sought isolation. After that other strategic factors could be weighed. The chosen site might be at a point controlling travel by road, at a central location on a private domain, on property stolen from the Church or even in a town or close to the wall of a fortified city. Beyond any such long-term considerations, however, the seigneur selected a setting that would enable him and his forces both to survive siege and to go over to the offensive effectively.

Under these circumstances it is not

MOTTE AT DOL in Brittany is attacked in a second detail. The light bridge at the left spans the motte's moat. A horse ridden by one of William's men is crossing the bridge to probe the motte's defenses. They are apparently weak: at the center Duke Conan of Brittany escapes by sliding down a rope to the far side of the moat. He fled to another motte at Dinan.

MOTTE AT DINAN surrenders in a third detail. The garrison continues to resist, but two of William's men have crossed the moat and are setting fire to the motte's timbers with torches. At the right Conan is seen extending his lance, passing the keys to William as a token of his submission. Notwithstanding this episode, only a few mottes were ever successfully stormed.

FORTRESS BOURG OF DAMPIERRE-LE-CHÂTEAU in the department of the Marne, depicted in this drawing after a 15th-century engraving, had grown from a motte-and-bailey castle between the 12th century and the time when the engraving was made. Only one structure remained on the summit of the motte. The associated bailey was surrounded both by a rampart and by a moat. Some of the bailey outbuildings, seen in the foreground, are on low ground. Others stand on higher ground with a guard tower at the right. Beyond the rampart and moat at the left are commercial, residential and religious structures, their occupants attracted by the stability of the bourg.



PLAN VIEW OF MOTTE AND BAILEY in the illustration at the top of the page shows its substantial dimensions: 120 by 60 meters at the base and 75 by 35 meters at the top. The motte earthwork itself was 15 meters high. The relation between the motte and the bailey is shown in a section along the line *A–B* that appears at the top left. It shows why that part of the bailey was known as the upper town.

surprising that a good many of the earliest *châteaux*, erected in unsatisfactory locations, were later abandoned by their builders. A number of anonymous vestiges still bear witness to the precariousness and instability of the first seigneurial habitations. The process of site selection, however, soon came to favor locations that were easy to defend, had a good view of the countryside and commanded terrain over which it was possible to venture in all directions.

What was *château* life like? Here is a 12th-century description of the dwelling tower of Ardres in French Flanders. The tower of this sizable *château* had three levels. At the ground level were kept the stores of grain and wine. On the next level was the seigneurial chamber, adjoined by smaller rooms occupied by the master's younger children, his table attendants (bearers of wine and bread) and some other servants. On the top level, which was also the level of the guard walk, were communal sleeping quarters for the master's older children, one for the sons and another for the daughters. Adjoining the dwelling tower was another three-story structure. There at the ground level meat was stored; on the next level were the kitchens and quarters for the cooks, and on the top level was the chapel, facing east. Actually few if any of the early *châteaux* were as spacious and comfortable as Ardres; a sensible fear of fire must have made most *château* builders place the kitchens farther away from the dwelling tower.

Nevertheless, at least at Ardres, the seigneur, his family, their table attendants and other servants all lived in the dwelling tower. Their cooks lived in the adjacent building and perhaps so did the chaplain and the sentries of the watch. The rest of the establishment—the knights of the garrison, the administrators of the seigneurial domain, artisans such as the blacksmith and others needed for the upkeep of arms and equipment, and lesser personnel such as stable boys—occupied buildings within the adjacent palisaded bailey. There stables, barns, ovens, workshops and similar outbuildings formed a small, self-contained community that existed only for the seigneur's service and under his protection. The small population of the typical early *château* scarcely interacted at all with the much more numerous peasants who inhabited the villages of the countryside.

This aloofness did not continue for long. By about A.D. 1050 clusters of outsiders began to establish themselves near the *châteaux*. The root causes of this shift appear to have been population growth and a resulting social instability. Carolingian Europe had been covered with deep forest, and in the 11th and 12th centuries forests were actively

A BERLITZ® LANGUAGE COURSE in a handsome BRIEFCASE for only $125 if *you* pay for it BUT wouldn't your company be interested in your speaking a second language?

These days a second language can help you help your company and so help *you* earn more. The many hispanic nationals now in our work force respond more readily to someone who speaks their language. Also, selling today is international. Whatever language it would pay you to speak, Berlitz has a do-it-yourself course in a handsome briefcase you can take anywhere. Contents—90 minute introductory tape, 40 lessons on five 60-minute cassettes, 6 illustrated word and phrase books, rotary verb finder to help you find verb tenses instantly. If not satisfied, return within 10 days and get your money back. Learn from the leader!

For your convenience on credit card orders dial toll-free
1-800-228-2028 ext. 35
and **refer to Dept. 3074**
24 hours a day, 7 days a week.

Why not give us a ring—right now! (our expense)

(Nebraska Residents Dial 402-571-4900)

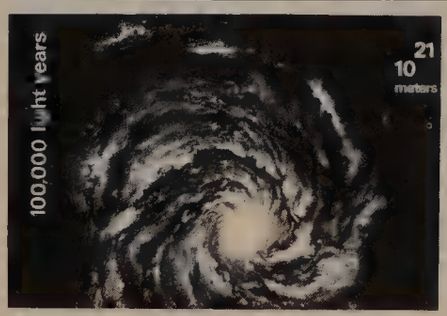
Berlitz Publications, Inc., Dept. 3074
3490 Lawson Blvd., Oceanside, NY 11572

Send **Berlitz Comprehensive Cassette Course(s)** checked. $125 each in briefcase plus $4 for shipping and insured delivery.

☐ **French** 86100 ☐ **German** 86101 ☐ **Italian** 86102 ☐ **Spanish** 86103

Enclosed check _____ money order _____ payable to Berlitz.
Or charge my AMEX _____ Diners Club _____ VISA _____
MasterCard Interbank # _____ Exp. Date _____
Card # _____ Exp. Date _____
Name _____
Address _____
City _____ State _____ Zip _____

N.Y. Residents add sales tax. Allow 4 weeks for delivery.



POWERS OF TEN
By Charles and Ray Eames

This famous film by Charles and Ray Eames transports you out to the edges of the universe and in to the micro-world of cells, molecules and atoms.

The film inspired the book POWERS OF TEN, the premier volume of the SCIENTIFIC AMERICAN LIBRARY.

POWERS OF TEN—another outstanding film available from PYRAMID FILM & VIDEO.

POWERS OF TEN 9 minutes, color
$225 Purchase: 3/4" U-Matic, 1/2" Beta II, 1/2" VHS, 16mm-Rent: $40 3/4" U-Matic and 16mm only
call toll free: 800-421-2304

PYRAMID FILM & VIDEO
Box 1048, Santa Monica,
California 90406-1048 (213) 828-7577

Discwasher Wants to Make You a Video Authority

For your copy of "THE VIDEO REFERENCE", complete the coupon below and return it to Discwasher, along with a check or money order for $1 to cover postage and handling.

Please send me _____ copies of "THE VIDEO REFERENCE". Enclosed is my check or money order for $1/copy to cover postage and handling.

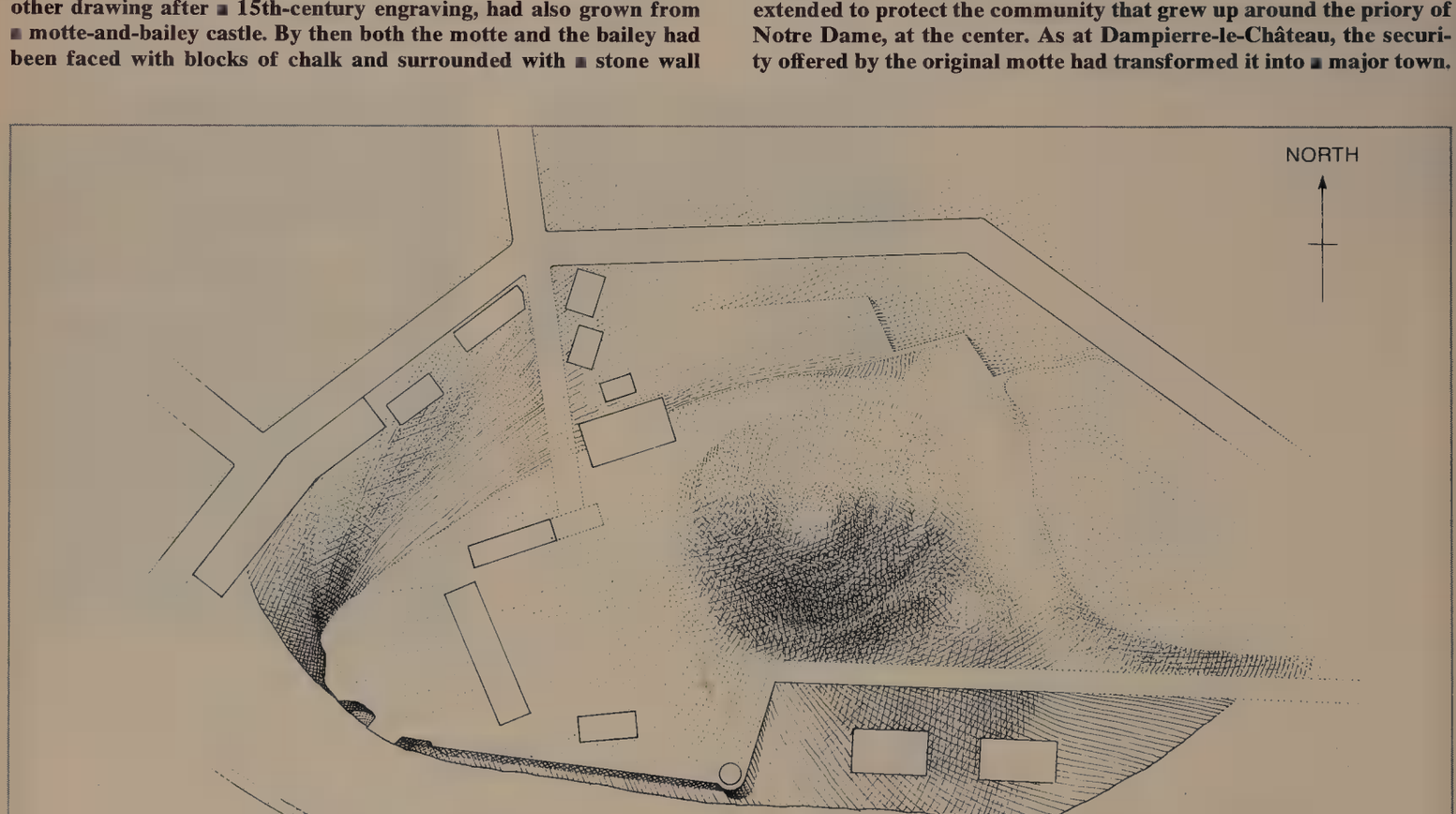
Name _____
Address _____
City _____ State _____ Zip _____

discwasher
1407 NORTH PROVIDENCE ROAD, P.O. BOX 6021, DEPT. SA, COLUMBIA, MO 65205 USA
A DIVISION OF JENSEN an ESMARK Company

cleared to allow agricultural expansion. In such newly cleared areas the interests of two quite different groups converged. The first group consisted of surplus labor hungry for land. The second was entrepreneurial: the younger sons of noble families, eager to set up their own domain. The convergence of interests led to the construction of a new kind of fortified site, where the seigneur's *château* and his followers' village all stood inside a single defensive earth wall. Examples are found in southern Aquitaine and Champagne.

Next, beginning in about A.D. 1070, local towns arose, particularly in the west of France. Churches and marketplaces were built in the immediate vicinity of one *château* or another. The population attracted to such amenities swelled the numbers in the growing set-

FORTRESS BOURG OF RETHEL in the Ardennes, depicted in another drawing after a 15th-century engraving, had also grown from a motte-and-bailey castle. By then both the motte and the bailey had been faced with blocks of chalk and surrounded with a stone wall that had towers placed at regular intervals. The same wall had been extended to protect the community that grew up around the priory of Notre Dame, at the center. As at Dampierre-le-Château, the security offered by the original motte had transformed it into a major town.



PLAN VIEW OF MOTTE shows how it took advantage of irregular terrain. The motte itself was on a rocky spur that extended eastward with an almost vertical drop on its south side and an only slightly shallower slope to the east and the north. The motte was 70 meters in diameter at the base and 10 meters in diameter at the summit. Probably built before A.D. 1026, it rose 21 meters above its base.

ComputerLand®

©ComputerLand 1983

No matter what you do, we can help you do it better.

All kinds of professionals are finding more than computers at ComputerLand. They're discovering new solutions to old problems and opportunities they hadn't imagined. Better ways to manage a practice, or a research project; to process a brief, or control a production process; to analyze an engine's performance, or a dancer's.

Because ComputerLand doesn't merely offer the widest selection of personal computers on earth. We have everything you need to use a personal computer to work more productively. Know-how to help you choose the right one in the first place. Plus software, service, accessories and professional advice to make the most of it.

Selection, support and expertise. Only ComputerLand provides them all—at over 400 stores around the world.

When it comes to personal computers, nobody does it better.

Why not come in and see what we can help *you* do better.

ComputerLand®
We know small computers. Let us introduce you.™

Over 400 stores worldwide.

For locations call 800-227-1617x118 (in California 800-772-3545x118; in Hawaii or Canada call 415-930-0777 collect).

tlement and soon the seigneur of the *château* found himself master of the adjacent bourg, or borough, economically and administratively the center of a sizable district.

Such bourgs were variants of other settlements, both in the open country and close to the few cities, that were multiplying everywhere at this time. They benefited from the same social dynamics and often enjoyed the same legal privileges. In any event villages without *châteaux* remained more numerous than villages with them. This suggests that a seigneur's effort to recruit a cadre of peasants in order to profit from the economic growth of the newly established village did not always conclude with a company of happy householders settled in the shadow of his tower.

When the fortuitous union of *château* and village gave birth to a bourg, the presence of the seigneur, his entourage and the other clientele of the same social class was certainly crucial to making the enterprise flourish. Yet for the bourg to become firmly rooted other factors favorable to growth had to exist. One such factor was the fertility of the local soil. Another was a location on a main road. Without such advantages there could be little probability of success. Negative factors also played a part; many a bourg withered away when its adjacent *château* was destroyed by chance or by malice.

These observations, concerned with the period of seigneurial consolidation from the middle of the 11th century into the 12th, have taken us a long way from the initial feudal period, a time of struggle when ordinary methods of defending habitations were shaky and when the new weapon, the motte, first came into service as a means of territorial domination. Let us now return to that earlier period in order to see the role the motte played in the evolution of a new kind of man and the molding of his behavior. In effect one cannot separate the feudal warrior from his environment.

To begin with, consider the contradictions of the times: on the one hand freedom and on the other a subjugated and exploited peasantry; on the one hand war and on the other a clergy preaching peace as the ideal. A seigneur inhabited a world of isolation and danger where—in order to protect himself and his family, to sleep in safety, to raise children to be knowledgeable about that harsh and hazardous life—he needed to hide behind an earth rampart and barricade himself on a motte. The only way to soften those harsh conditions was to depend on the links of kinship or, better, on the parent-surrogate links of vassalage and an oath of fealty.

Even with those assurances constant vigilance was necessary. At home the seigneur lived in fear of treason. In battle the knowledge that the stakes were high sharpened cunning and incited treachery. Even a task as simple as finding a wife became enmeshed in haggling or led to matchmaking that left little room for personal choice.

Then, little by little, the violence and insecurity of these early feudal times gave rise, as it were, to their own antibodies. The fear of treason served in the end to reinforce the fidelity of the vassal and to exalt the loyalty of the warrior. The art of war, learned through continuous training from an early age, was not sidetracked into becoming the art of ambush. Instead it evolved into straightforward prowess, so that a victor faced by a valiant but unlucky adversary—virtually a mirror of himself—was careful to exercise fair play. By the same token the young warrior's lust, stimulated by the presence of both unmarried and married women within the *château,* was transformed into civility.

This kind of courtesy was not limited to the men's attitude toward women. It defined the deportment of the men in the seigneur's retinue. Within the confines of the *château* the choice was to live either in fellowship or in antagonism. Clearly the best course was to adhere to certain rules that made life as pleasant as possible or, at the least, to refrain from violence. Thus it was that in the confined world behind the palisade there slowly developed the moral standards that, when feudal society entered its second age at the end of the 11th century, inspired the troubadours. They sang of chivalry and love, but what the songs actually celebrated was the dual achievement of stabilization and colonization. Those foremost figures of chivalry, the seigneur's retinue of knights, began as simple warriors who had attained a higher estate through the practice of cavalry warfare. At the same time their garrison duties imposed on them specific values of courtesy and honor: the rules of the game.

One can conclude that the motte of the 10th and 11th centuries has a historical significance that has not attracted enough notice. As an instrument of domination par excellence it should be evaluated in terms of technological efficiency. The waterwheel mill and the draft-animal harness played a comparable technological role in the economic revolution of the Middle Ages. Consider, however, another factor. The authority of Louis XIV, who ruled France for more than half of the 17th century and on into the 18th, is regarded as having been nothing less than absolute. The fact remains that the Sun King in all his glory was not as well obeyed as any minister in the French republic of today. His orders reached the farthest boundary of his realm, but they got there slowly, and the functionaries on the spot were left a certain margin for interpretation. It is not so today. For example, it is enough for the Minister of the Interior to book a conference call, to lift his telephone and have some 100 departmental prefects standing at attention at the other end of the line. Moreover, any ambiguity in his orders can be eliminated as he goes along. As a means of exercising authority in government the telephone not only is faster than the written order but also can carry more weight.

Viewed in proportion and at a somewhat different level the motte played a role similar to the telephone's but in reverse. It diminished the king's prerogatives, affording the seigneurs a means of weakening central authority and giving birth to as many seigneuries as there were *châteaux.* Unable to exercise sovereignty, the king could only act the part of a feudal overlord, hoping for obedience no longer from subjects but rather from vassals. The conclusion is self-evident: however powerful the ruler and however determined the unruly, the outcome for both remained moot as long as neither had an effective technological advantage.

A second conclusion is that the motte certainly affected the mass of the rural population. From having been more or less self-sufficient they found themselves tied to the *châteaux* and their seigneurs and forced to pay taxes to these new overlords. The effect on the behavior of the overlords, however, was even more profound. Accustomed to a life in large open dwellings in a countryside populated with peasant freemen, the seigneurs in a matter of decades found themselves huddling behind ramparts, armed to the teeth in order first to grasp power and then to keep it. Clinging to this technological instrument of terror, the motte, the seigneurs and their retinues began to evolve a culture both primitive and original. On the one hand the culture was two-faced and dedicated to naked power; on the other it was a culture of courage and faithfulness to the given word. Certain of the popular medieval ballads, the chansons de geste, captured this evolving culture as if in a living mirror.

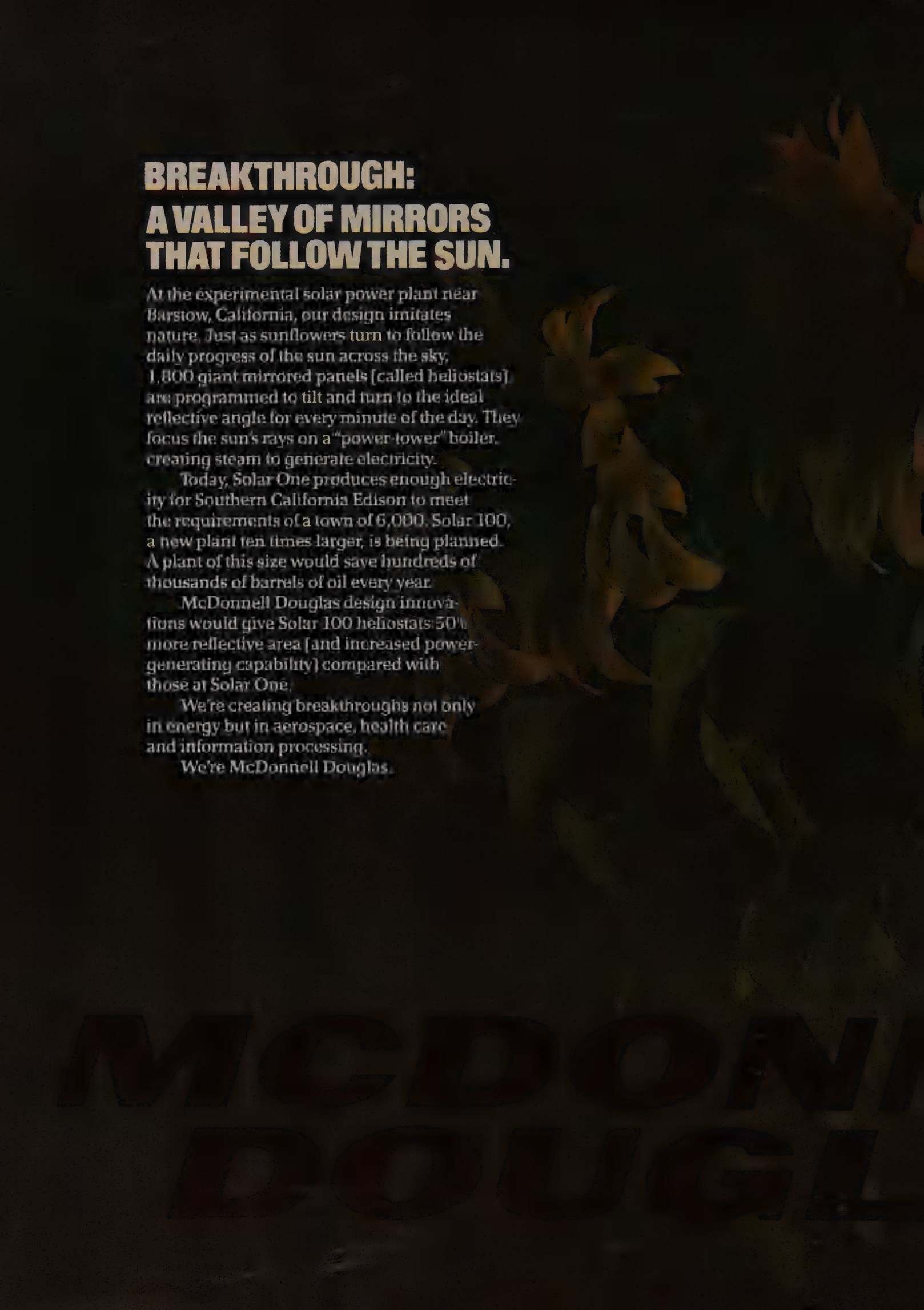
Little by little, as the generations passed, a certain equilibrium arose. A set of new values acknowledged seigneurial mutuality and made it possible to live less at sword's point. The *châteaux* became open to friends and neighbors; tournaments took the place of warfare; shields, once anonymous, carried symbols indicating the identity of the bearer. Where there had reigned ruse and brutality there now flourished virtues such as prowess, generosity, courtesy and honor. Such is the legacy the period bequeathed us from its second stage of development, a legacy, bred in the environment of the motte, that deserves to be called the civilization of the *châteaux.*

A ripple never stops.

It goes all the way to the edge of the water. And then comes back. Every time you reach out from yourself, you make a bright new one. U.S. Swimming and U.S. Diving are a chance for thousands of young people to find that you can reach even farther than you thought. And make ripples of your own. And every time you make a ripple, it touches everyone around you.

PHILLIPS 66

Phillips Petroleum has sponsored U.S. Swimming for eleven years, and U.S. Diving for five, helping with operating costs of the national and international competitions that lead to the Olympics. We continue to support them. Because their ideals touch the lives of everyone.



BREAKTHROUGH: A VALLEY OF MIRRORS THAT FOLLOW THE SUN.

At the experimental solar power plant near Barstow, California, our design imitates nature. Just as sunflowers turn to follow the daily progress of the sun across the sky, 1,800 giant mirrored panels (called heliostats) are programmed to tilt and turn to the ideal reflective angle for every minute of the day. They focus the sun's rays on a "power-tower" boiler, creating steam to generate electricity.

Today, Solar One produces enough electricity for Southern California Edison to meet the requirements of a town of 6,000. Solar 100, a new plant ten times larger, is being planned. A plant of this size would save hundreds of thousands of barrels of oil every year.

McDonnell Douglas design innovations would give Solar 100 heliostats 50% more reflective area (and increased power-generating capability) compared with those at Solar One.

We're creating breakthroughs not only in energy but in aerospace, health care and information processing.

We're McDonnell Douglas.

MCDONNELL DOUGLAS

Sudden Cardiac Death: A Problem in Topology

Many sudden deaths are the result of fibrillation: a disruption of the coordinated contraction of heart muscle fibers. The cause may lie in a state of affairs described by a mathematical theorem

by Arthur T. Winfree

On the campus of the McGill University School of Medicine in Montreal the morning of November 7, 1914, was bright and chilly. It was a Saturday, and there were few professors or students around to notice George Ralph Mines enter his laboratory to begin some weekend experimental work. Mines, a 28-year-old physiologist, had been doing work on fibrillation, a radical disorganization of the heart's pattern of contraction. Ordinarily the heart contracts as a unified mass, but in fibrillation each small region of cardiac muscle contracts rapidly and with no apparent coordination with the neighboring tissue.

Mines had been trying to determine whether relatively small, brief electrical stimuli can cause fibrillation. For this work he had constructed a device to deliver electrical impulses to the heart with a magnitude and timing that could be precisely controlled. The device had been employed in preliminary work with animals. When Mines decided it was time to begin work with human beings, he chose the most readily available experimental subject: himself. At about six o'clock that evening a janitor, thinking it was unusually quiet in the laboratory, entered the room. Mines was lying under the laboratory bench surrounded by twisted electrical equipment. A broken mechanism was attached to his chest over the heart and a piece of apparatus nearby was still recording the faltering heartbeat. He died without recovering consciousness.

The death of George Mines might have remained a personal tragedy of little scientific interest if it had not been for the subject of his last work. Fibrillation is one of the major causes of the syndrome known as sudden cardiac death, which kills several hundred thousand people per year in the U.S. Moreover, of the several causes of sudden cardiac death fibrillation is the least understood. It can arise with little warning in apparently healthy people. In many instances when an autopsy is done, there is no perceptible clinical reason for the drastic interruption of the normal coordination of the heart. Mines's work shed light on part of the problem. He was the first to demonstrate that fibrillation can develop after a relatively small electrical impulse if the impulse is applied to the heart at the right time. Although much work has been done on sudden cardiac death and fibrillation since 1914, it is still not known conclusively how a small stimulus can elicit such a drastic response in the heart.

Intriguingly, the branch of mathematics known as topology can help to elucidate the problem. Like certain other physiological systems, cardiac tissue is capable of a rhythmic electrical discharge. If an external electrical impulse is delivered to a piece of tissue from the heart, the most frequent effect is to shift the usual rhythm ahead or back, without changing the interval between impulses after the stimulus. By means of a topological theorem, however, it can be shown that there must be a relatively small electrical stimulus that does not yield such a predictable shift if it is applied at the right moment in the cycle of beats. Indeed, the experimental results indicate that after such a singular stimulus the regular beating can stop.

The cessation of contractions in a small region of heart muscle is not fibrillation. It can be shown topologically, however, that around a patch of tissue where the beat has stopped the conditions are favorable for the creation of a circulating wave of electrical impulses. Such a wave could circulate through the heart, usurping the function of the normal pacemaking apparatus. In passing, the circulating wave could give rise to many smaller waves, causing small areas of the heart to contract rapidly in an uncoordinated way. The best [illegible] t models suggest that a patte[illegible]- lating waves is present i[illegible]

The arguments I shall present here describing how a singular stimulus can cause fibrillation have by no means been proved. They are primarily mathematical arguments and their confirmation will require much practical work in physiology. Even if they are confirmed, their clinical implications are far from clear. Nevertheless, topology has provided a significant new insight into the problem of sudden cardiac death.

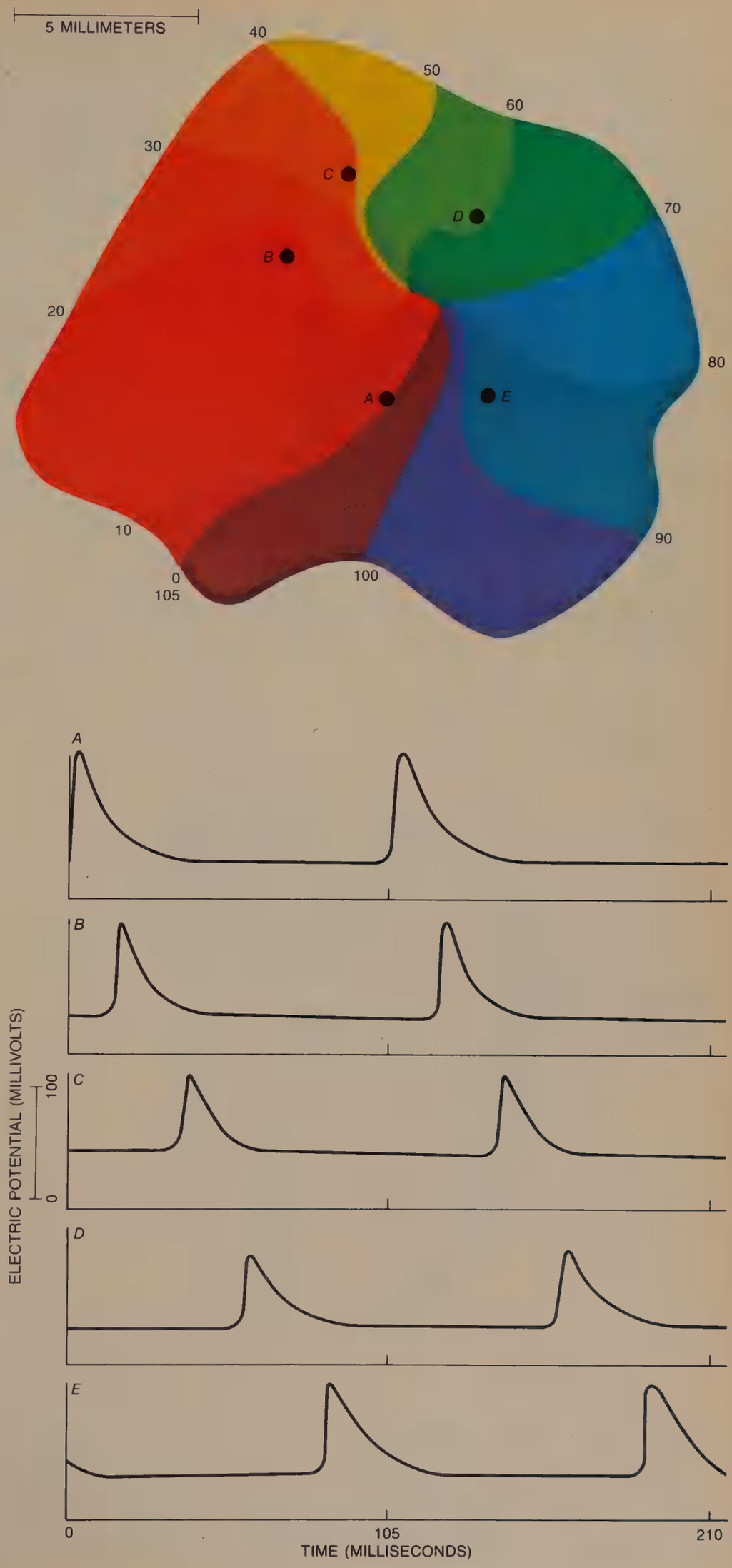
Before topology can be applied to heartbeat timing we must briefly consider the normal structure and function of the mammalian heart. The heart is an extraordinarily complex organ and a complete description of it is not possible here. The following much-simplified description concentrates on the properties of the heart that are pertinent to the topological argument. The heart consists of four chambers. The two atriums at the top of the heart serve as priming chambers. The larger ventricles below them are the main pumping chambers. Although in its normal operation the heart contracts more or less as a single mass, it can be noted when the action is observed closely that the contraction begins in the atriums and soon thereafter spreads to the ventricles.

The coordination of the four chambers is maintained by the passage through the cardiac tissue of a wave of electrical impulses that travels quickly from cell to cell. The heart is composed mainly of several types of specialized muscle cells that have the form of long fibers. When a fiber is in the resting state, the electric potential inside the cell membrane is somewhat lower (more negative) than the one outside the membrane. The stimulus for each fiber to contract is a rapid increase in the potential inside the membrane, at the end of which the interior of the cell briefly becomes electrically positive with respect to the environment before the potential decreases again.

The changes in electric potential are mediated by the flow of several ions (atoms with a net electric charge) across the cell's outer membrane. The rapid increase in potential is called a depolarization, and it can propagate along the muscle fiber as an "action potential" much like the ones that convey information in the nervous system. After the passage of an action potential there is a refractory period of a few tenths of a second during which normal stimulation cannot elicit another action potential in the fiber.

The fibers in the heart are closely apposed to each other. In addition specialized organelles in the cell membrane of each fiber provide regions of low electrical resistance. These organelles make it possible for the triggering electrochemical impulse to pass from cell to cell. Therefore an impulse originating in one part of the heart can be transmitted quickly through the entire organ. Moreover, certain tissues in the heart have the capacity to spontaneously generate triggering impulses at regular intervals. The regions are called pacemakers and the one with the highest intrinsic rate imposes its rhythm on the entire heart.

The fastest spontaneous firings in the healthy heart come from a small mass of tissue near the top of the atriums known as the sinus node. The sinus node directly initiates the contraction of the atriums. Its impulse also propagates to the atrioventricular node, a second pacemaker lying between the right atrium and the right ventricle. From the atrioventricular node the impulse is relayed to the muscular wall of the ventricle by the Purkinje fibers. Purkinje fibers are adapted for very rapid conduction of electrical impulses. They extend throughout the ventricles in an arrangement that resembles the roots of a tree. When the pacemaking impulse passes from the Purkinje fibers into the ventricular muscle, a wave of contrac-



A ROTATING WAVE of electrical impulses can circulate indefinitely in the heart and could underlie fibrillation. In a fibrillating heart small regions of tissue contract rapidly and without coordination. If fibrillation persists for more than five minutes, the victim usually dies. The trigger for heart muscle cells to contract is an electrochemical impulse. Under certain conditions the impulse can travel in a circular path. The upper panel shows the path of a rotating wave induced experimentally in rabbit heart tissue by Maurits Allessie and his colleagues at the Center for Biomedicine at Maastricht in the Netherlands. Colored areas indicate the distance traveled by the wave front in 10 milliseconds; each complete rotation required 105 milliseconds. The traces in the lower panel show the electric potential at the five labeled points in the tissue. The periodic sharp increases in potential mark the passage of the rotating wave front.

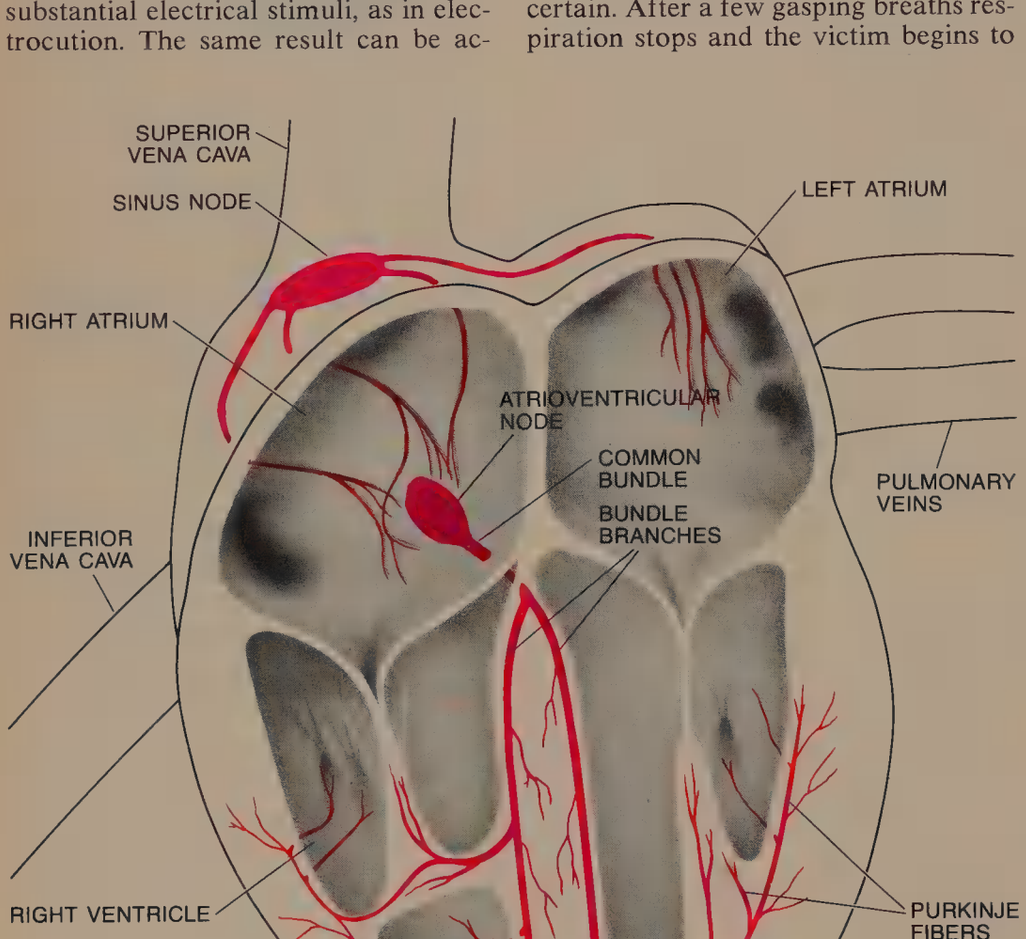
tion spreads through the ventricles in milliseconds.

At rest the human heart contracts roughly once per second. The intrinsic beat in the sinus node that determines the resting rate can be accelerated or retarded by impulses carried by nerves coming from the brain, various ganglia and the internal organs. Such nerves infiltrate the entire heart, but their endings are particularly dense in the pacemaking system. For example, the vagus nerve can have a considerable effect on local rates of contraction and on the coordination of the entire heart.

The acceleration and retardation of the heart's contractions are usually smoothly synchronized. The mechanism can fail, however, if there is an infarction: an interruption of the flow of blood to a portion of the heart. Synchronization can also fail if the heart is exposed to unusually high levels of hormones or ions. Even without physical damage or chemical stress synchronization can fail if the heart is perturbed by substantial electrical stimuli, as in electrocution. The same result can be accomplished by a small stimulus if it is unfortunately timed. We shall see how such small stimuli can cause sudden cardiac death.

The term sudden cardiac death was coined in 1887 by John A. McWilliam of the University of Aberdeen, although the phenomenon had been known long before then. McWilliam noted that in some instances of sudden cardiac death the heart did not simply stop but entered into violent and disorganized activity. In the 19th century such violent activity was referred to as *delirium cordis,* or madness of the heart. The modern name, fibrillation, was introduced in the 1870's. According to Willis A. Tacker, Jr., and Leslie A. Geddes of Purdue University: "When a fibrillating heart is held in the hand, it feels like a wad of writhing worms. In many cases the rate of this random process is so rapid that the heart surface seems to shimmer. In other instances multiple waves of contraction and relaxation are clearly visible." If fibrillation persists for more than five minutes, death becomes almost certain. After a few gasping breaths respiration stops and the victim begins to turn blue because of a lack of oxygen in the blood. Convulsions and groaning sometimes accompany the stopping of respiration.

There is little agreement among physiologists on the mechanism of fibrillation. Studies of the membrane in single cells have been the source of much information about how electrical impulses are relayed in physiological systems. Such work has been particularly informative about the cells of the nervous system. Work with single cells, however, has been of only limited help in understanding fibrillation. The reason is that fibrillation is a disorder of the organization of the heartbeat rather than a malfunction of the individual fibers: it affects the timing of activity in the heart fibers with respect to one another. Indeed, it is probable that in the fibrillating heart each fiber is responding in the usual way, but the pattern of electrical impulses has been disturbed, resulting in a loss of synchronization.



HUMAN HEART is ■ four-chambered pump. The atriums are priming chambers. The ventricles are the main pumping chambers. The heart's contractions are synchronized by electrical impulses originating in the sinus node. The node, which has ■ regular intrinsic electrical discharge, serves ■■ the pacemaker for the heart as ■ whole. Its impulse initiates the contraction in the atriums directly. The impulse is also relayed to the atrioventricular node, and from there it is conveyed to the muscular ventricles by specialized conducting fibers, among them the Purkinje fibers. Under ordinary conditions the heart contracts almost ■■ ■ single mass.

What could cause the thousands of fibers in the human heart to suddenly lose the precise coordination they have maintained for a lifetime? Maurits Allessie of the Center for Biomedicine at Maastricht in the Netherlands, pointed out to me that in the paper Mines submitted to a medical journal just before his death he proposed that fibrillation could be caused by an electrical wave circulating in the heart muscle.

In his work Mines delivered electrical impulses to the heart, varying the timing of the stimulus to cover the interval between contractions. Most of the stimuli Mines applied had no lasting effect, and none initiated an unmistakable circulating wave in the heart tissue. He noted, however, that "under some conditions a stimulus of very brief duration may induce fibrillation . . . if properly timed. . . . The stimulus employed would never cause fibrillation unless it was properly timed." The precise moment between beats of the heart when a small stimulus can cause fibrillation was rediscovered many years after Mines's death; it is now known as the vulnerable phase.

My own interest in the vulnerable phase was aroused partly by my being a helpless witness to two sudden cardiac deaths. At the time I was working on the biological mechanisms of circadian rhythms, the intrinsic 24-hour cycles many organisms exhibit. Certain topological features of circadian timing systems had led me to predict that a single brief stimulus, if it was properly timed, could cause organisms that had been quite rhythmic to lose their regular rhythm. I had found such stimuli and demonstrated their effect, as other workers have done since.

It turns out that the topological properties of circadian clocks that make them susceptible to disruption by a sin-

gle stimulus are shared by physiological systems capable of regular electrical discharges, such as nerve cells, heart muscle fibers and even the heart as a whole. A topological property is a property of a geometric figure that persists when the figure is bent, twisted, stretched or otherwise deformed in a continuous manner. For example, an image seen through a distorting lens is topologically equivalent to the undistorted image.

Topology might be defined as the study of properties that remain unchanged in spite of quantitative changes (but not qualitative ones). Inferences based on topological features of a physical system can be quite effective in making predictions without reference to the precise quantitative aspects of the system. Topological reasoning was valuable in my work on circadian rhythms because little was known about their mechanism, and so quantitative models depended largely on guesswork. In contrast, so much quantitative information has been gathered on the coordination of the heartbeat that the assembled data can be overwhelming. By confining one's attention to certain topological properties of heartbeat timing it may be possible to cut through the mass of data and reach an understanding of how fibrillation begins.

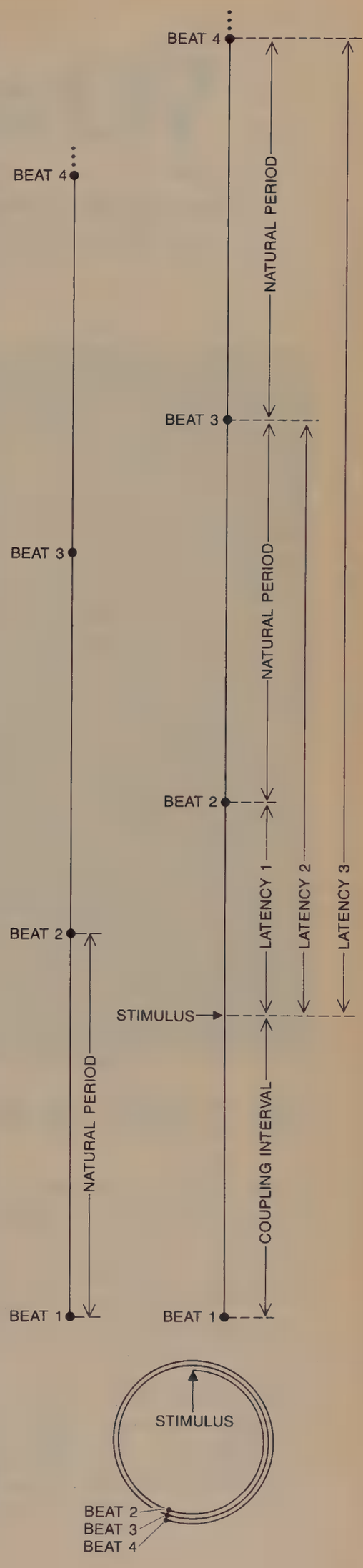
The topological qualities of the heartbeat that are of interest here were first suggested by observations made in about 1930 by J. C. Eccles of the University of Oxford. Like Mines, Eccles was examining the effect of the timing of an external stimulus on the subsequent contractions of the heart. Eccles, however, was concerned with the timing of the next few beats rather than with circulating waves or fibrillation. He stimulated the heart of the dog by activating the vagus nerve, which, as I noted above, is one of the regulators of the rate of contraction in the normal operation of the heart. Eccles observed that after a stimulus the heartbeat was perturbed but quickly returned to its usual rhythm. In many instances the heart recovered so rapidly that every complete cycle after the stimulus had the normal length; the effect of the stimulus was merely to offset the usual rhythm.

Since 1930 a vocabulary has been developed to describe such changes in timing. The period from the initial beat to the application of the stimulus is called the coupling interval. The period from the stimulus to a subsequent beat is called the latency. Since there are many beats after the stimulus, there are many latencies: each successive latency is the period until a new beat. (It should be noted that the topological principles under discussion can be applied to several biological systems. In this context "beat" does not necessarily refer to the action of the heart. The beat is simply the event that marks the beginning and end of each cycle in the system under observation.)

Depending on the length of the coupling interval and the magnitude of the stimulus, the first beat after the stimulus can be either delayed or advanced. Both Mines and Eccles observed the effect of applying a stimulus with a constant magnitude at progressively later times in the cycle. When the coupling interval is varied in this way, the coupling interval and the latency can show either of two distinct relations. Which relation is observed depends on the magnitude of the stimulus.

One of the two patterns is called weak rescheduling. To understand weak rescheduling it is valuable to consider what happens when the stimulus is too small to affect the timing of the subsequent beat. Under these circumstances the sum of the coupling interval and the latency until the first beat after the stimulus must equal the natural period, or the interval between beats in the unperturbed system. Hence the coupling interval and the latency must vary inversely: as the stimulus is given progressively later in the cycle, the latency becomes progressively shorter. On a graph that relates latency to coupling interval all beats lie on a diagonal that slopes down to the right. As the coupling interval is increased smoothly from zero to the full natural period, the latency decreases smoothly over one full cycle. As a result every possible value of latency from zero to one full period appears once.

It is characteristic of weak rescheduling that every value of latency appears an odd number of times when the coupling interval is varied over its full range. The pattern of later beats result-



ELECTRICAL STIMULUS applied to the heart can advance or delay the next beat. In the unperturbed cycle (*vertical line at upper left*) the interval between beats is called the natural period. When an external stimulus is given to the heart (*vertical line at upper right*), the interval from the initial beat to the stimulus is called the coupling interval. The interval from the stimulus to a subsequent beat is called the latency. Since there are many beats after the stimulus, there are many latencies. Any complete natural periods in the latency can be removed by winding the line standing for the latency around a circle with a circumference equal to the natural period (*circle at bottom*). When this is done, the points standing for beats after the stimulus lie on top of one another. These later beats have all been advanced or delayed by the same interval and therefore successive beats are one natural period apart. The main effect of the stimulus is to shift the usual rhythm ahead or behind; the natural periodicity is maintained. The terminology employed to describe such shifts can be applied to any system that is capable of rhythmic firing; "beat" then simply means the event that marks the end of the cycle.

What if you chose as a technical

"At Union Carbide, an HP computer network automates production management and has paid for itself in less than two years."

Union Carbide's Linde Division in Tonawanda, N.Y., manufactures custom production equipment. Factory systems manager Doyce Coffman says, "When production demand fluctuates, it complicates our scheduling, product costing, and long range planning.

"To solve these problems we automated production and materials management with a network of HP 1000 and 3000 computers and 43 data capture terminals.

"Because these systems continuously monitor operations and respond in real time, we can identify and eliminate production bottlenecks on line. The payback? Improved efficiency factory-wide and a dollar savings that covered the cost of our entire HP network in less than two years."

What should you expect from a computer partner?

Quality—from researching customer needs to product development, manufacturing, marketing, after-sales service, and support!

Quality, by HP's definition, encompasses more than the product.

This high quality recently was recognized by the Union of Japanese Scientists & Engineers, who awarded a coveted Deming Prize to Yokogawa-Hewlett-Packard, HP's joint venture company in Japan.

YHP offers the entire spectrum of HP products and in 1981 accounted for Japanese sales of more than $200 million. Nearly 60 percent of the products YHP sells are imported from HP manufacturing operations

The Deming Prize: *for outstanding merit in the control of quality.*

Hewlett-Packard computer partner?

"At Likoff Cardiovascular Institute, an HP computer processes electrocardiograms in minutes instead of hours."

Likoff Cardiovascular Institute, in Philadelphia, is a leading center for the research and treatment of cardiovascular disease. Using an HP computerized ECG management system, the Institute processes 100,000 electrocardiograms a year, 70,000 from outside facilities.

Dr. Gary Anderson, heart station director, says, "By automating the ECG processing cycle, we reduced turnaround time per test from 2½ hours to just 4 minutes—and reduced patient costs by 10%.

"With the HP Management System, we can retrieve patient records instantly, including often critical comparison data in our evaluations. And because this system handles incoming ECGs on a priority basis, we're responding to emergencies much faster than before, with better interpreted ECGs at a lower cost."

the US and abroad. Conversely, HP manufactures and exports arious products to HP customers roughout the world.

ne total quality control approach r which YHP was awarded the eming Prize is a point of pride roughout HP. Customers in every rt of the world should expect *ality* as a fundamental element in computer partnership with HP.

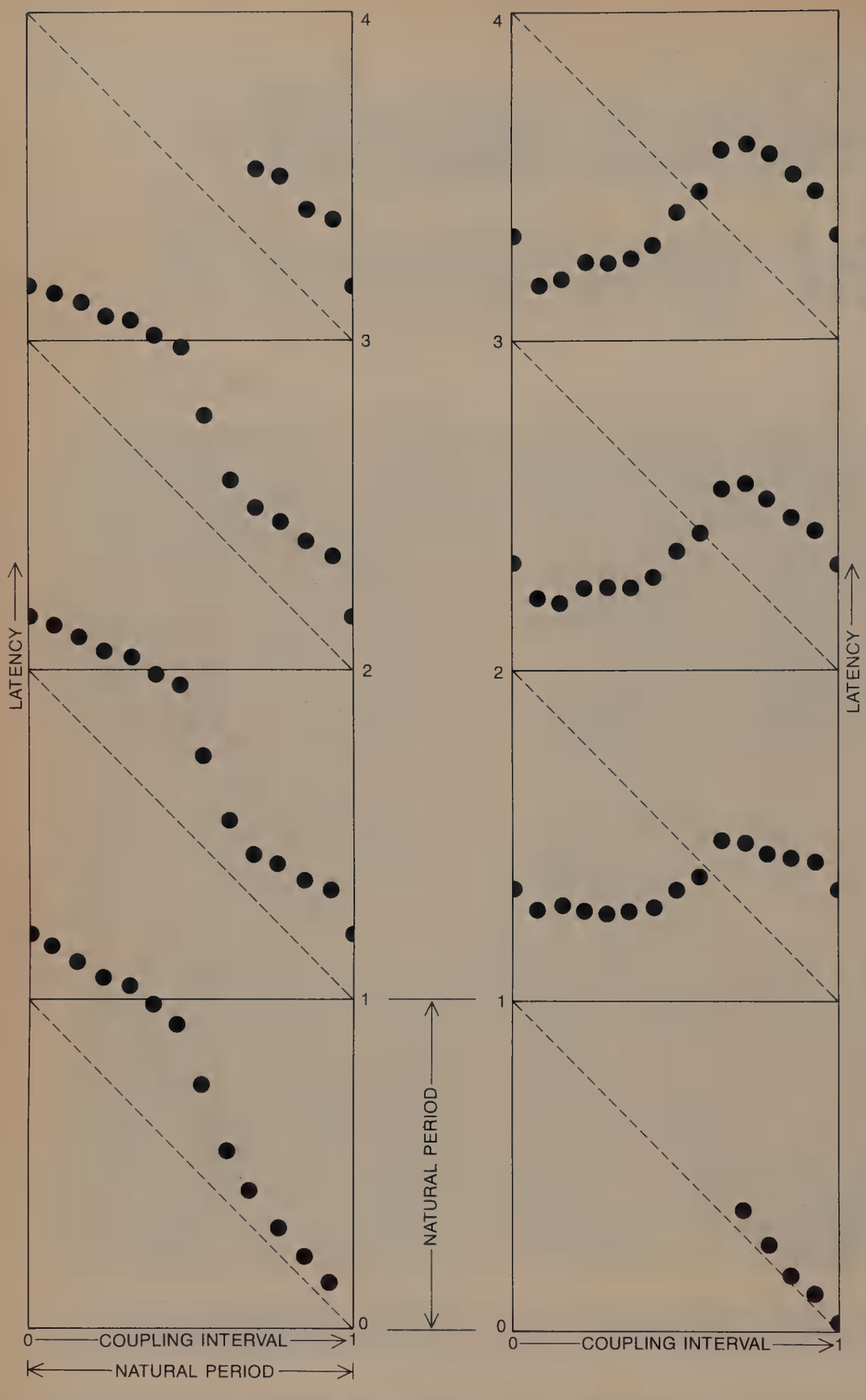
You should consider a working partnership with HP. Now.

For a free copy of our brochure explaining HP's practical, proven approach to meeting your long-term computer and information needs, write to A.P. Oliverio, Senior Vice President, Marketing, Dept. 207, Hewlett-Packard Company, P.O. Box 10301, Palo Alto, CA 94303.

When performance must be measured by results

HEWLETT PACKARD

0002341



WEAK AND STRONG RESCHEDULING are two topologically distinct patterns of beats following the application of an external stimulus. The data in the graphs come from work with rabbit hearts done by Jose Jalife and Joseph J. Salata at the Upstate Medical Center of the State University of New York at Syracuse. In each graph the dots stand for the beats that result when a stimulus with a particular strength is applied at a coupling interval varying from zero to one period. If the stimulus were too small to advance or delay the next beat, the coupling interval and the latency would vary inversely and the dots would fall on the diagonal in each square. If the stimulus is sufficiently strong, the next beat is advanced or delayed and the dots do not fall on the diagonal. If the strength of the stimulus is below a certain level, however, the latency still decreases smoothly over one natural period as the coupling interval is increased. This pattern is called weak rescheduling (*left*). If the stimulus is increased beyond a certain level, the latency undergoes no net increase or decrease as the coupling interval is varied through one period. The latency can vary by any amount as the coupling interval is increased, but it returns to the initial level when the coupling interval reaches one full period. This pattern is called strong rescheduling (*right*). In the example shown the latency decreases to a minimum, then increases to a maximum before returning to the initial value. That the two patterns are topologically distinct can be shown by joining the left and right edges of each graph to form a vertical tube. (Since the cycle repeats itself every period, the edges are equivalent.) The dots on the left graph then form a continuous helix and the dots on the right graph form three separate rings.

ing from a negligible stimulus is merely the limiting case of weak rescheduling. If the magnitude of the stimulus is increased, the next beat can be either advanced or delayed, depending on the timing of the stimulus. If the strength of the stimulus is not increased beyond a certain level, however, the latency still decreases smoothly and continuously as the coupling interval is increased through one full cycle. The two quantities are no longer inversely proportional; the plot of latency can curve above and below the diagonal as the coupling interval is increased. Nevertheless, each value of latency still appears an odd number of times.

If the strength of the stimulus is increased above a certain level, the second pattern, strong rescheduling, appears. In strong rescheduling the latency undergoes no net increase or decrease as the coupling interval is increased over one full period. The latency can increase or decrease by any amount as the coupling interval is varied, but it returns to the initial value when the coupling interval reaches one complete natural period. For example, when the coupling interval is zero (meaning that the stimulus coincides with a natural beat), the latency could be about one and a half natural periods. When the coupling interval is increased, so that the stimulus comes progressively later in the cycle, the latency decreases to a minimum, then increases to a maximum and ultimately returns to the initial value as the coupling interval approaches the natural period. As the coupling interval is varied through the complete cycle not all values of latency need appear. Furthermore, the values that do appear are seen twice: once as the latency is decreasing and once as it is increasing.

The curve described above is only one possible example of strong rescheduling. Many other curves are possible. What all such curves have in common is that each value of latency that appears does so an even number of times. Most values do not appear; the ones that do can appear any even number of times. The range of values of the latency that do not appear widens with increasing stimulus strength. In the limiting case of strong rescheduling only one value of the latency appears. The limiting case follows a stimulus so strong that it yields the same result regardless of when in the cycle it is applied; the next beat always comes, say, half a second after the stimulus.

In both weak and strong rescheduling the pacemaker resumes its normal periodicity after the advanced or delayed second beat. Weak rescheduling has long been known. Almost every rhythmic biological process shows weak rescheduling if the stimulus is sufficiently small. The topologically distinct strong

THE CAR:

New Renault Alliance. We set out to build the best small car in the world.

Affordable European technology. Our goal was an alliance of technology and affordability. Our achievement is Alliance, a fine European sedan–built in America–for the price of a four-passenger econobox.

Comfort for five. Alliance has eight inches more rear seat hip room than Escort. To comfortably ride five, not four. And with the front seats on pedestals, rear seat passengers can slide their feet forward and enjoy the ride.

Inspired performance. Front-wheel drive. Fully independent suspension. Power front disc brakes. Rack and pinion steering. Electronic ignition. All are standard.

Alliance's 1.4 litre engine is electronically fuel injected. Small wonder, then, that Alliance produces such outstanding mileage.**

52 EST HWY | 37 EPA EST MPG**

Alliance 2-door

$5595*

Built in America. To last. Alliance is covered by American Motors' exclusive Buyer Protection Plan.® With the only full 12-month, 12,000-mile warranty that covers virtually every part even if it just wears out. Plus a five-year No Rust-Thru Limited Warranty.™

Renault Alliance. The Alliance of quality and affordability is here. At Renault and American Motors dealers.

RENAULT
American Motors

THE AWARDS:

Motor Trend experts name Alliance 1983 Car of the Year.

Finalists were judged for styling and design, quality, comfort, ride and drive, handling, fuel economy, performance and dollar value.

"The Renault Alliance is the best blend of innovation, economy, and fun-to-drive we have seen in almost a decade. Moreover, it represents a uniquely successful blend of outstanding European engineering and American manufacturing know-how."
–Tony Swan, Editor
Motor Trend Magazine

And the European version of Alliance was named Europe's 1982 Car of the Year by 52 journalists from 16 countries.

Alliance DL 4-door

FETY IS A SNAP

*Manufacturer's suggested retail price for the Alliance 2-door. Price does not include tax, license, destination charges and other optional or regionally required equipment. **Compare 1983 EPA estimates for the Alliance 2-door with estimated MPG for other cars. Your actual mileage depends on speed, trip length and weather. Actual highway mileage and CA figures will probably be lower.

Over 300 manufacturers sell their microcomputers with Microsoft software. There's a reason.

Powerful simplicity. That's the concept behind every software program we write. Powerful programs that let you spend more time thinking about the problem...and less time thinking about the computer. That concept has earned us the confidence of over 300 microcomputer manufacturers. Starting with the very first microcomputer you could buy. And today, Microsoft® languages, operating systems and applications software are running on well over a million microcomputers. Worldwide.

Made for each other. Manufacturers literally build their computers around Microsoft software. That's no exaggeration. In fact, Microsoft frequently participates in the initial design and development of microcomputers. That's a major reason why Microsoft software runs so well on the majority of the world's 8- and 16-bit systems.

User-oriented. Microsoft is the only software supplier to offer a full range of compatible operating systems, languages, utilities and applications programs. If you're not a computer expert, here's what that means to you: Better programs. Programs that are not only more powerful, but easier to learn and use.

Better tools. We can't buy your trust. We have to earn it. With better products. Tools you can easily use to solve complex problems. We started with the first BASIC for the first microcomputer. Today, we offer a broad range of proven tools for microcomputers. Including SoftCard™ and IBM® RAMCard,™ products that materially increase the capabilities of Apple ® and IBM PC computers. Plus, business and management software such as Multiplan,™ the powerful electronic worksheet you can learn to use in just a few hours. Better tools. Because they're designed specifically for your computer.

Ask your Microsoft dealer. Most microcomputer manufacturers offer their systems with some Microsoft software, but you'll undoubtedly want more. Programs and languages that solve your specific problems. Programs for your industry, business or home. Programs that do more, yet ask less. Ask your computer software dealer for a demonstration. You'll see why more than 300 microcomputer manufacturers offer their systems with Microsoft software. The reason is powerful simplicity.

BETTER TOOLS FOR MICROCOMPUTERS

MICROSOFT™

MICROSOFT CORPORATION
10700 NORTHUP WAY
BELLEVUE, WASHINGTON 98004

Microsoft is ■ registered trademark, and Multiplan, SoftCard, RAMCard and the Microsoft logo are trademarks of Microsoft Corporation. Apple and the Apple logo are registered trademarks of Apple Computer, Inc. IBM and the IBM logo are registered trademarks of International Business Machines Corporation. The Radio Shack logo is ■ registered trademark of Radio Shack, a Division of Tandy Corporation. The Burroughs logo is ■ registered trademark of Burroughs Corporation. The Victor logo is a registered trademark of Victor Business Products. The Wang logo is a registered trademark of Wang Laboratories, Inc.

rescheduling, however, was recognized in neural pacemakers only five years ago. The data on strong rescheduling had been published some time before, but it took almost a decade for the distinct pattern to be recognized. The existence of strong rescheduling in neural pacemakers has now been confirmed by many workers, most recently by Jose Jalife and his student Joseph J. Salata of the Upstate Medical Center of the State University of New York at Syracuse. Jalife and Salata found both weak and strong rescheduling by stimulating the sinoatrial node of the rabbit and examining the subsequent impulses of the pacemaker.

As we shall see, the fact that the pacemaking apparatus of the heart is capable of strong rescheduling provides a crucial link in the argument leading from a single small stimulus to fibrillation. To understand how this could be so it is necessary to consider the latency in detail. As noted above, for each stimulus there are many latencies, corresponding to the many beats that follow the stimulus. For the purposes of topological analysis it is advantageous to subtract from each latency any complete natural periods included in it. There is a scheme for visualizing the result of this operation that leads directly to the topological argument.

Imagine a ring with a circumference equal to the natural period of the pacemaker. The ring is marked with a full cycle of colors arranged in the order of the spectrum that just fits around the circumference. Since the circumference of the ring is equal to the natural period, each color can be made to stand for a particular fraction of the period. For example, if red stands for a negligibly short latency, then yellow stands for a latency of about a third of the period, blue a latency two-thirds of the period and red again a latency one full natural period long.

Visualize the latency following a particular stimulus as being represented by a long, infinitesimally thin cord. One end of the cord corresponds to the time of application of the stimulus. A short distance away is the point corresponding to the first beat after the stimulus. The points marking the subsequent beats fall at intervals along the cord. The stimulus end of the cord is held against the red section of the colored ring and the rest of the cord is wrapped around the ring as many times as it will go.

When this procedure is carried out, the points representing beats after the stimulus lie on top of one another. Their position on the ring indicates the fundamental latency, that is, the time elapsed from the stimulus to each subsequent beat, minus any whole number of natural periods that can be subtracted. If the points happen to lie in the red part

QUESTAR® . . . *the telescope that can take it*

A Questar's durability is the thing that amazes so many owners. That an instrument so small and so beautiful can exhibit such toughness in the clinches is the quality that astonishes almost as much as its unique optical performance. Hubert Entrop, who sends us those great deep-sky photographs, keeps writing to tell us how he sometimes gets his best shots of galaxies and nebulas in areas that are exposed to sudden mountain squalls and gusting winds, but his Questar keeps guiding steadily along. Now comes a letter from *Wings*, a group that specializes in safaris to far away places. Will Russell, their leader, writes:

"WINGS now owns five Questars and during the last two years we've used them at Pt. Barrow, Alaska, in Kathmandu and Jerusalem, along the Amazon, in Kenya, Siberia, Thailand and many places in between. We've used them in the steaming tropics and sub-zero cold; in blizzards, dust storms, Atlantic hurricanes, Pacific typhoons and just about every other sort of weather. We've dropped them, kicked them over, watched them blow over, had them partially dismantled by foreign security guards, run over by airlines luggage conveyors and temporarily confiscated. In short, in the past 24 months we've learned quite a lot about using Questars."

And so, whether you propose to use your Questar on your patio or in one of those distant spots where Russell's Questars have been put through their paces, you will be reassured to know that Questars can take it. We like to think that everyone babies his or her Questar, but obviously Wings isn't able to do this and the instruments not only survive but turn in fine performances.

It is nice to know, too, that this old, established company still guarantees its product for ten years against defective parts and workmanship. Even Questars that have fallen down a flight of stairs or were dropped out second story windows, have turned up here to be put back together for a whole new useful life.

© 1982 Questar Corporation

Our booklet describing Questar, the world's finest, most versatile telescope, contains an astonishing collection of photographs by Questar owners. Send $2 for mailing costs on this continent; by air to S. America, $3.50; Europe and N. Africa, $4.00; elsewhere, $4.50.

QUESTAR

Box C, Dept. 202, New Hope, Pa. 18938

Medicine is always changing, now there's a text that changes right along with it.

SCIENTIFIC AMERICAN *Medicine.*

The answers that you get from SCIENTIFIC AMERICAN *Medicine* are current, because each month new chapters replace the old. You will know where to look for the answers, because this advanced text follows the familiar structure of medical knowledge, changing as medicine changes.

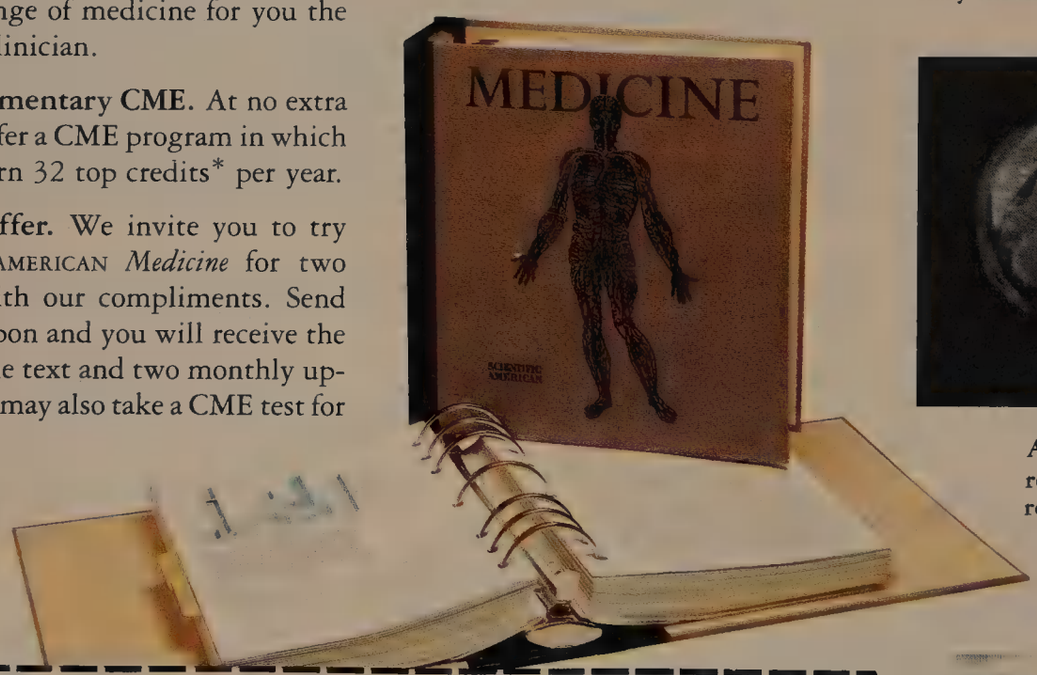
Comprehensive. Leading authorities from Harvard and Stanford cover the full range of medicine for you the seasoned clinician.

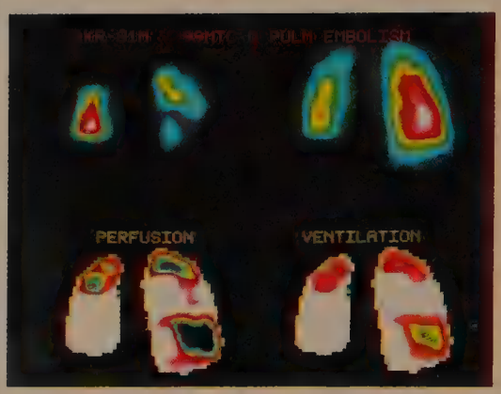
Complimentary CME. At no extra cost, we offer a CME program in which you can earn 32 top credits* per year.

Trial Offer. We invite you to try SCIENTIFIC AMERICAN *Medicine* for two months with our compliments. Send us the coupon and you will receive the two-volume text and two monthly updates. You may also take a CME test for credit. At the end of 60 days, if you decide to continue the subscription, you will owe us $220 for the first 12 months (current renewal is $170); otherwise, return the books; there is no charge.

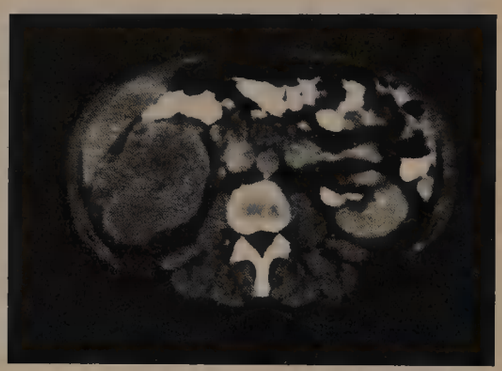
Please mail the coupon today so that we can put medical progress to work for you.

You may also order your trial copy by calling toll-free 1-800-345-8112 (in Pennsylvania, 1-800-662-2444).

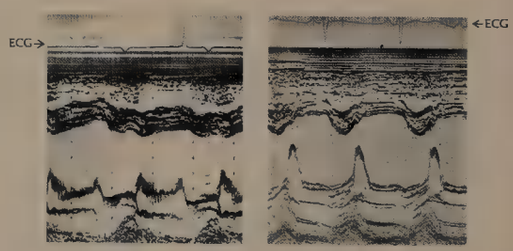




Computerized scintigraphy reveals pulmonary thromboembolism.



Abdominal computed tomogram reveals large renal carcinoma replacing part of right kidney.



Echocardiograms from patients with aortic regurgitation.

SCIENTIFIC AMERICAN MEDICINE

415 Madison Avenue, New York, New York 10017

Please enroll me as a subscriber to SCIENTIFIC AMERICAN *Medicine*. On receipt of this coupon you will send me the advanced two-volume text described in your announcement and update it regularly by sending me new monthly subsections. I understand that the annual subscription of $220 for SCIENTIFIC AMERICAN *Medicine* is tax deductible. If I am not entirely satisfied, I may cancel at any time during the first 60 days, returning all materials for a *complete refund*.

☐ Please enter my subscription for SCIENTIFIC AMERICAN *Medicine*
☐ I shall also enroll in the CME program
☐ I enclose a check made out to SCIENTIFIC AMERICAN *Medicine* for $220*
☐ Bill me ☐ VISA ☐ MasterCard

Expiration Date_____ Account Number_____

*Please add sales tax for California, Illinois, Massachusetts, Michigan and New York

Name_____ MD Specialty_____

Address_____

City_____ State_____ Zip_____

Signature_____

Please allow six to eight weeks for delivery. All payments must be in U.S. dollars. CME is available outside the U.S. and its possessions for an additional $65. Surface routes used unless airmail delivery is prepaid.

7B

*As an organization accredited for continuing medical education, the Stanford University School of Medicine designates this continuing medical education activity as meeting the criteria for 32 credit hours in Category 1 for Educational Materials for the Physician's Recognition Award of the American Medical Association, provided it has been completed according to instructions.

This program has been reviewed and is acceptable for 32 prescribed hours by the American Academy of Family Physicians.

This program has been approved by the American College of Emergency Physicians for 32 hours of ACEP Category 1 credit.

The Hughes story continues...

Engineering Story.

Begins with you. Your talent. Your skills and Hughes Support Systems.

Hughes Aircraft Company. A pioneer in high technology electronics.

Hughes Support Systems provides a broad and deep spectrum of support activities including: field engineering, logistics support, training of customer personnel on Hughes' systems, technical manuals and other training materials. Product lines include computerized training simulators and a wide variety of automatic test systems.

Hughes support engineering capabilities encompass programs ranging from electro-optics, lasers and missiles to airborne radar systems. Hughes creates the electronics — Support Systems shows the world how to use them.

This diverse organization created the F/A-18 Simulator. Electronically, this device uses a combination of state-of-the-art disciplines to recreate the sights, sounds and sensations of flight.

With 1,500 different projects company-wide, Hughes offers challenge and technological growth.

Our current openings include:

- Analog/Digital Circuit Design
- Automatic Test Systems Design
- Calibration
- Electro-Optical Testing
- Electro-Optics Circuit Design
- Engineering Writing
- Field Engineering
- Logistics Support Analysis
- Maintainability
- Product Support
- Project Engineering
- Real-Time Programming
- Software Engineering
- Systems Engineering
- Training Engineering

If you'd like to be part of a story that's still unfolding, write yourself in. Send your resume to:

Hughes Support Systems
A1/1B404, Dept. 950
P.O. Box 9399
Long Beach, CA 90810-0463

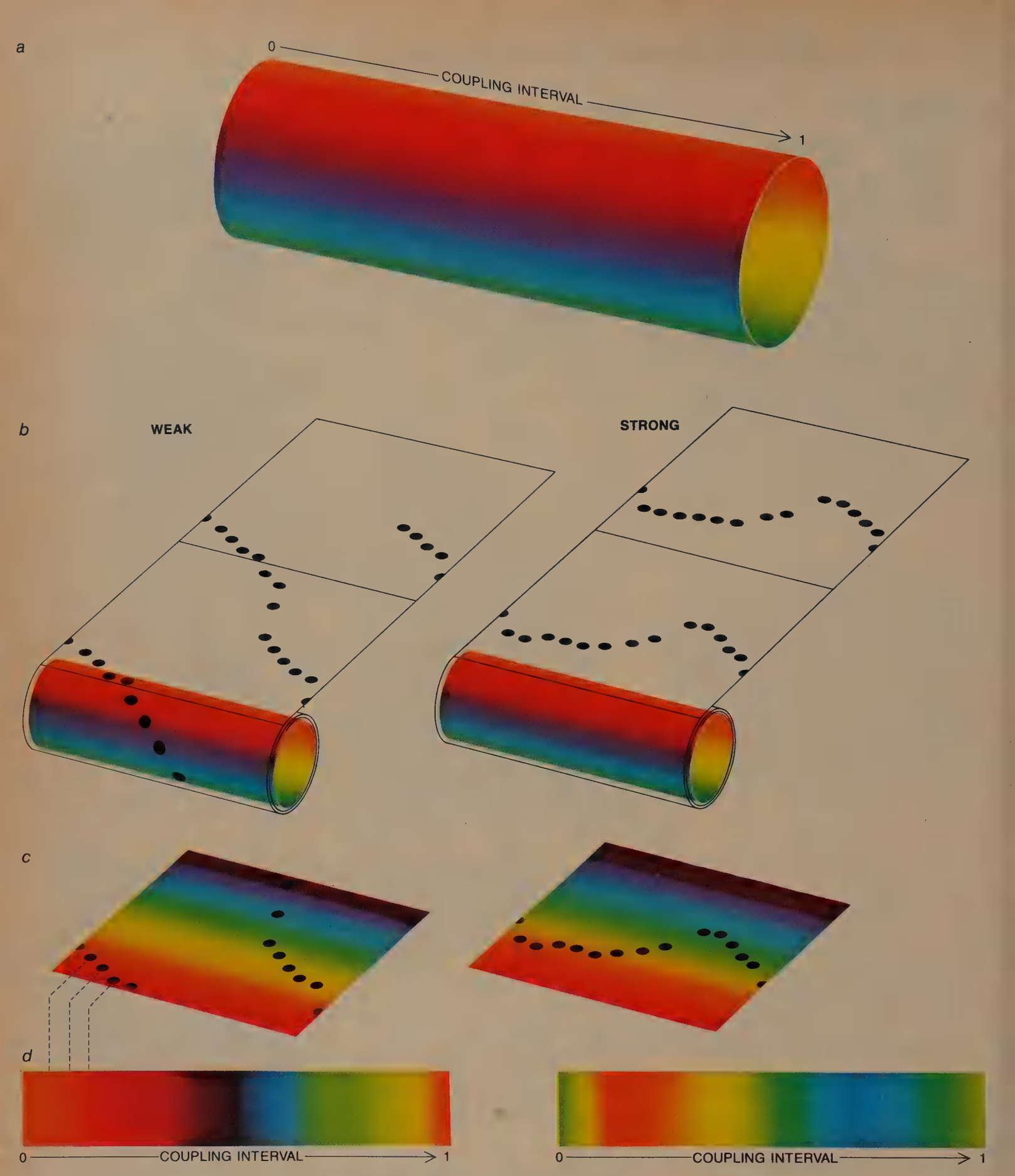
Proof of U.S. Citizenship Required/Equal Opportunity Employer

Creating a new world with electronics

HUGHES
HUGHES AIRCRAFT COMPANY

Support Systems

Write yourself in.



LATENCIES CAN BE ASSIGNED COLORS by means of a cylinder with a full cycle of colors around its circumference; the colors are arranged in the order of the spectrum (*a*). The circumference of the cylinder equals one natural period. Therefore each hue can represent a fraction of the natural period. Red stands for a latency of zero or one full period. The graphs of weak and strong rescheduling shown in the illustration on page 148 are placed against the cylinder with the bottom edge of each graph (equivalent to a latency of zero) resting on the red band. The graphs are then wrapped around the cylinder; each square panel of the graph wraps exactly once around the circumference (*b*). In weak and strong rescheduling successive beats after the stimulus occur one period apart. Hence after the wrapping successive rows in each graph lie on top of one another. The procedure removes any complete natural periods from the latency. The cylinder with the graph wrapped around it is slit open along the edge of the red band and laid flat, yielding a row of dots on a square panel (*c*). The row of dots represents the latencies that result when a particular stimulus is applied with a coupling interval varying over the natural period. The colors on which the dots lie are arranged horizontally according to the coupling interval (*d*). The sequence at the left shows weak rescheduling. The sequence at the right shows strong rescheduling. If the stimulus were too small to affect the timing of the next beat, the wrapping would yield a horizontal sequence beginning with red and running smoothly through the full cycle of colors back to red.

of the ring, the fundamental latency is equal to zero or the natural period of the system. If the points lie in the yellow part, the fundamental latency is about a third of the natural period. Because the colors of the ring correspond to fractions of the natural period, the latency has been assigned a color that stands for its length.

Any stimulus that yields weak or strong rescheduling merely offsets the natural cycle of the pacemaker. Any such stimulus, depending on its timing, can therefore be associated with a fundamental latency and hence with a color on the ring of latencies. Now consider what happens when the colors that stand for latencies are put in the appropriate places on a rectangular graph with coordinates representing the coupling interval and the strength of the stimulus.

The coupling interval is shown along the horizontal axis at the bottom of the rectangle, with zero at the left and one full period at the right. The strength of the stimulus is shown along the vertical axis at the left edge of the rectangle, with zero at the bottom. The points along the lower edge of the rectangle represent the result of applying a negligible stimulus at a coupling interval ranging from zero to one full period. As I have indicated, such a stimulus yields latencies that decrease smoothly through one natural period. Therefore when each point along the lower edge of the rectangle is assigned a color according to its fundamental latency, a full cycle of colors in sequence appears.

Along the upper edge are the points that stand for the result of applying a very strong stimulus at a coupling interval that varies over the full cycle; these are the conditions leading to strong rescheduling. As we have seen, strong rescheduling can result in a latency that begins at a value of about one and a half natural periods, decreases to a minimum and then increases to a maximum before returning to the initial level. Thus the upper edge of the rectangular graph could be colored in a sequence that includes green at the left, orange, green and blue in the middle and green again at the right.

The colors along the left and right edges of the rectangle represent the result of applying a progressively stronger stimulus at a coupling interval of zero (at the left) and of one full period (at the right). There could be many different sequences of colors along the vertical edges, but as we shall see the exact arrangement is inconsequential from a topological point of view. It should be noted, however, that the pattern must be the same on both edges. The reason is that a coupling interval of zero is equivalent to a coupling interval of one full period, and so stimuli given at those intervals must generate the same pattern of latencies.

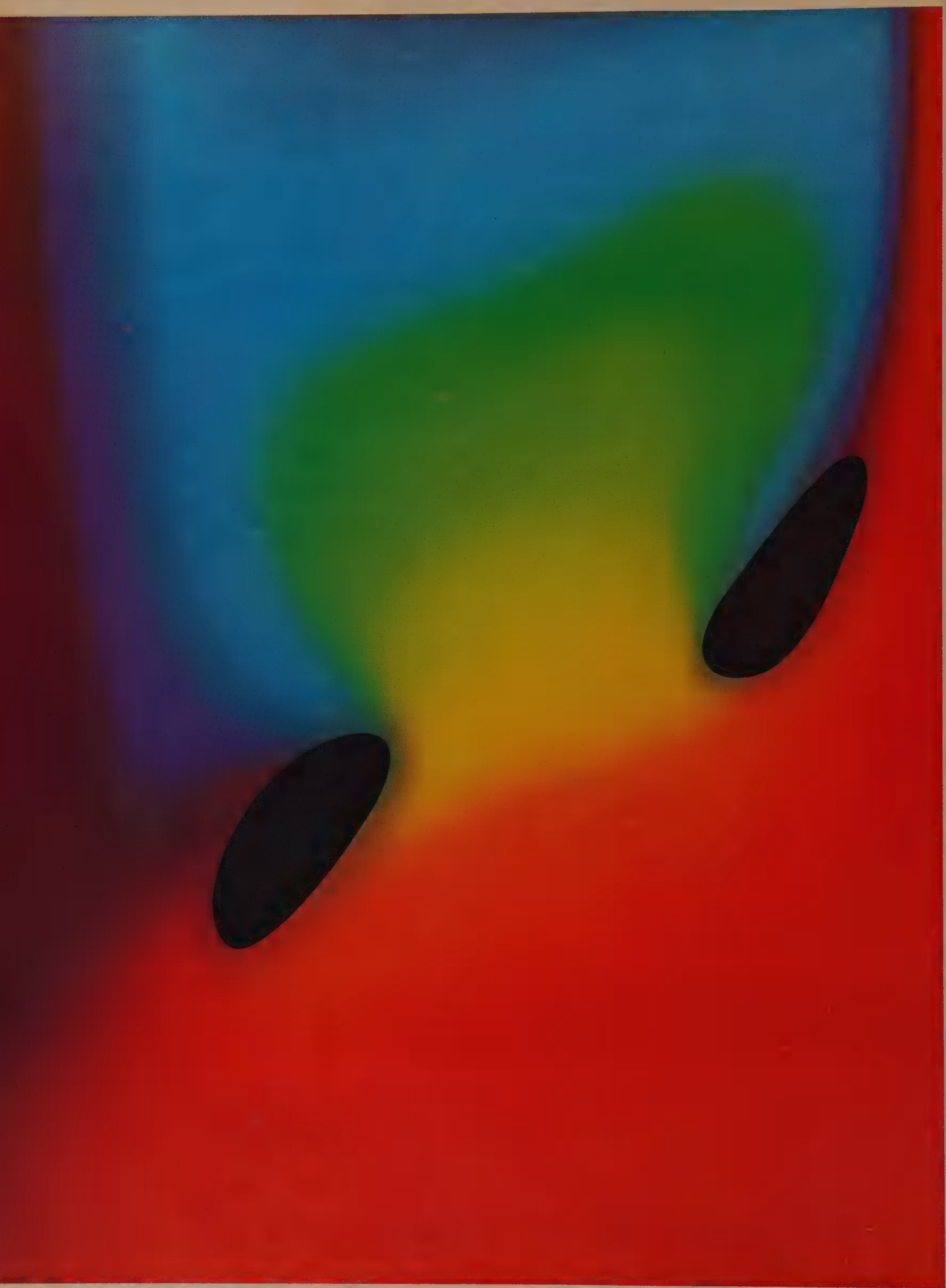
RECTANGULAR GRAPH can be partially filled in with colors that stand for latencies. Each point on the graph stands for a specific combination of coupling interval and stimulus strength. The color at the point designates the resulting latency. The lower edge of the rectangle includes the full cycle of colors that stands for the effect of applying a negligible stimulus at a coupling interval varying from zero to one period. The left and right edges show the effect of applying a progressively stronger stimulus at a coupling interval of 0 (at the left) or 1 (at the right). The upper edge represents strong rescheduling. Moving around the perimeter of the rectangle, one cycle of colors from red back to red is traversed. Beginning at the lower left corner and moving right, the full cycle of colors is encountered along the bottom of the graph. Up the right edge is a smooth sequence from red to green. Across the top is a sequence from green through blue, green, orange and back to green. Down the left edge is a smooth sequence from green to red: the same sequence as on the right edge but in the reverse order. Thus there is one full cycle of colors on the perimeter of the rectangle. When the perimeter has such a cycle, a circle can be inscribed somewhere in the rectangle on which the colors of the cycle appear once each and in sequence. A topological theorem called the nonretraction theorem shows that it is impossible to associate every point inside the circle with a point on the circle itself and also preserve the original continuity of the points within the circle. Therefore the theorem implies there is at least one point in the interior of the graph that cannot be associated with a color. In other words, a particular combination of stimulus strength and coupling interval does not yield a consistent latency. Such a stimulus represents a mathematical singularity. Its effect on the regular rhythm cannot be predicted, but it will not result in a mere offset of the pacemaker's cycle.

The edges of the rectangular graph have now been colored. To continue the visual exercise imagine a wheel much like the ring of latencies, with a full cycle of colors applied to its rim. The wheel is placed at the lower left corner of the graph and then moved all the way around the perimeter of the rectangle, beginning by moving to the right along the lower edge. As the wheel is moved it is turned so that the color on the wheel matches the color appearing underneath it on the graph.

What net advance will the wheel have made by the time the circuit is complete? Along the lower edge of the graph the wheel makes one complete forward revolution as it traverses the latencies corresponding to a negligible stimulus. Going up the right edge, the wheel turns from red to green. Along the upper edge the wheel turns from green through blue, green and orange before returning to green; hence the net forward rotation is zero. Going down the left edge of the rectangle, the pattern of colors on the right edge is encountered again but in reverse order. Therefore, whatever the effect of going up the right edge was, it is reversed by going down the left edge.

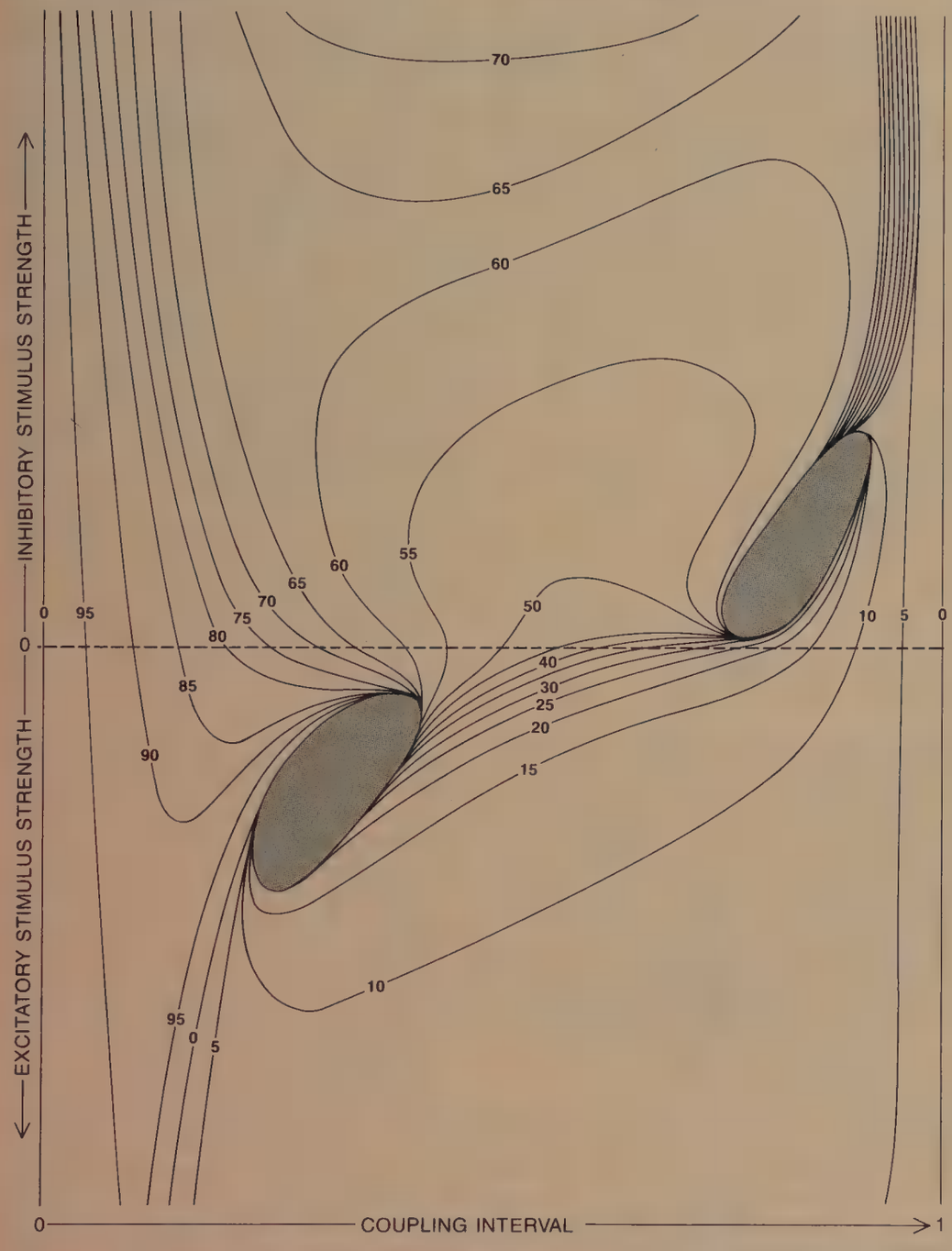
It is apparent that the upper edge of the rectangle and the vertical edges make no ultimate contribution to the turning of the wheel: their turns are self-canceling. When the wheel returns to the red area at the lower left corner of the rectangle, the only contribution to its turning that has not been reversed is the contribution of the lower edge of the graph. Thus the wheel makes one complete forward revolution as it traverses the perimeter of the graph of latencies.

What about the points in the interior of the rectangle? One might expect to find that every possible combination of coupling interval and stimulus strength is associated with some definite fundamental latency. It would then follow



that every point in the rectangle could be assigned a color. Furthermore, one would expect the latency to change continuously with any small change in coupling interval or stimulus strength, so that the colors in the interior of the rectangle would flow smoothly into one another without discontinuities. It is the surprising prediction of topology that such a smooth coloring of every point is not possible.

The topological theorem called the nonretraction theorem shows that it is impossible to completely retract the points on the surface of a disk onto the boundary of the disk and at the same time ensure that if two points are adjacent on the surface, they will also be adjacent when they are retracted onto the boundary. Only if at least one point is not retracted can the rest be rectracted while maintaining their original continuity. A physical example of the implications of the nonretraction theorem is the fact that a thin film of soap in a circular holder cannot shrink onto the holder unless the film is pricked. The topological significance of the pricking is that it removes one point from the disk, allowing all the other points to find appropriate places on the boundary and maintain their continuity. When the singular point is removed, the rest of the film can be continuously retracted onto the holder.



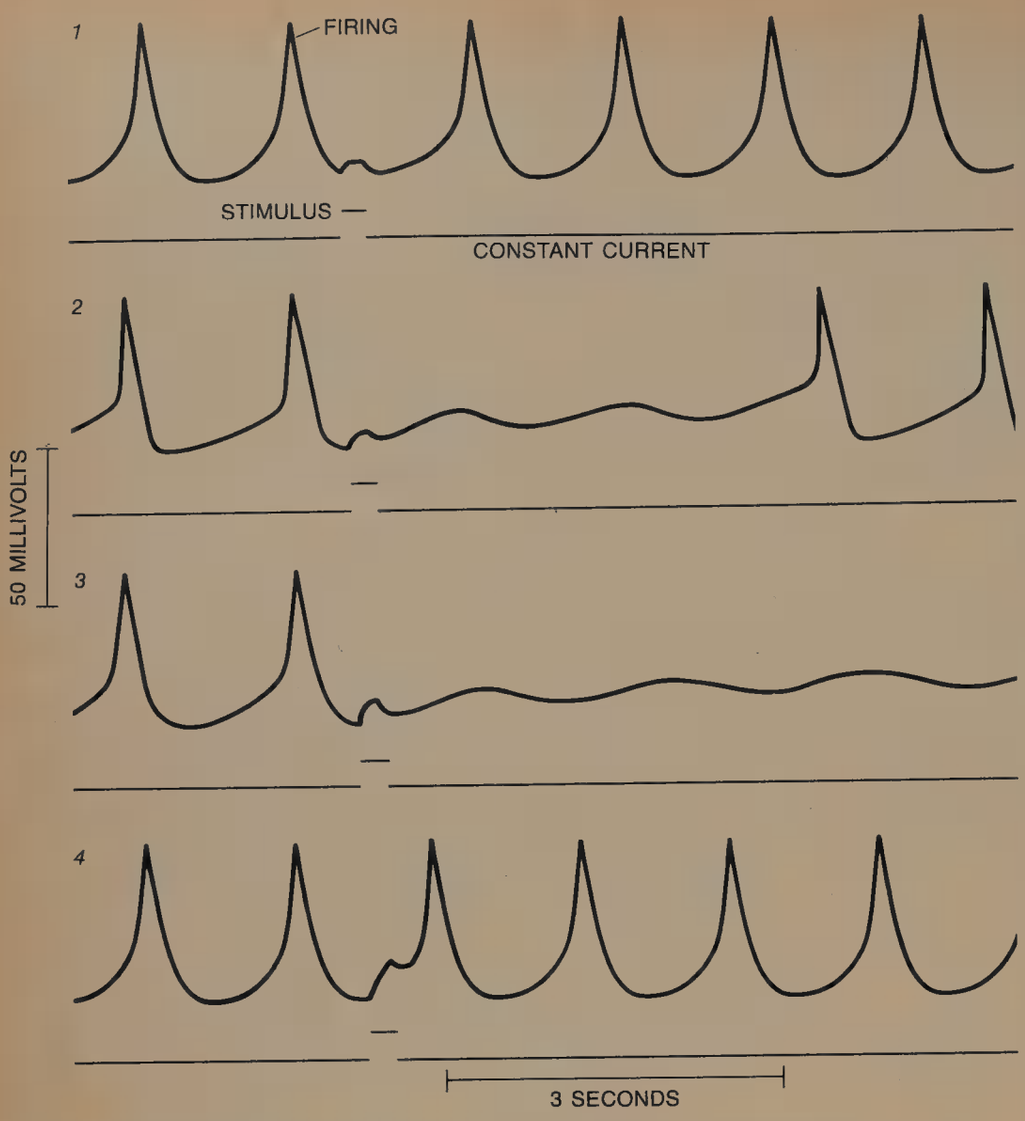
"BLACK HOLES" corresponding to singular stimuli (the ones that do not yield consistent latencies) were found by the author's student Eric Best in a computer simulation of the rhythmic firing of the squid giant axon. A mathematical model was employed to simulate the response of the axon to various stimuli. The stimulus strength and the coupling interval were varied, and the resulting latencies were assigned colors. The squid axon is sensitive to both excitatory and inhibitory stimuli. Therefore the simulation yielded two rectangles like the one in the illustration on page 153. The two rectangles can be joined along the line that represents a stimulus of zero strength. The result is the colored graph at the left. The black area in each rectangle represents stimuli after which the axon does not resume rhythmic firing. Surrounding each black area is a full cycle of colors in sequence according to the duration of the latency. The colors stream in an orderly way from one black area to the other. When points on the graph that have the same latency are connected, the result is a set of contour lines like the ones on a topographical map. The contours from the squid-axon simulation are shown above. When such contours represent latencies, they are called isochrons. The number on the isochron indicates a percentage of the natural period. The topology of the colored graph would be unchanged if the graph were laid out on a region of the heart. If this were done, the isochrons would represent the latencies of the cells in the region of tissue, that is, the interval after which each cell would fire.

As we saw in the exercise with the wheel, the perimeter of the rectangular graph includes a complete cycle of colors. Some colors, however, appear more than once on the perimeter. Green, for example, appears at least four times: once on the lower edge and three times along the upper edge. It can be shown that if there is a full forward cycle on the perimeter of a rectangle, it is possible to inscribe in the interior of the rectangle a circle on which the colors of the cycle appear once each and in order. The nonretraction theorem implies that all the points inside the circle cannot be retracted onto the circle and also maintain their original continuity. It is possible to imagine the points inside the circle flowing smoothly toward the circle itself, perhaps in a pattern where each point moves toward the nearest point on the circle. No matter what the pattern of flow, however, there must be some point whose neighbors diverge: that point cannot itself find a place on the circle.

If some point inside the circle cannot be associated with a point on the circle, it follows that there is some point inside the circle (and hence inside the rectangle) that cannot be given a color. Since each point in the rectangular graph stands for a particular combination of stimulus strength and coupling interval and each color stands for a fundamental latency, the nonretraction theorem implies there is some stimulus that, if it is given at a particular time in the cycle, does not result in a fundamental latency.

The point in the rectangle that cannot be colored is called a singular point, and the corresponding combination of stimulus strength and coupling interval is called a singular stimulus. The singular stimulus will not be particularly large, since it falls somewhere within the rectangle. The upper edge of the rectangle stands for stimuli that yield strong rescheduling and the lower edge for stimuli that yield weak rescheduling. Therefore the singular stimulus must have a strength between the strengths of the stimuli that yield the two kinds of rescheduling.

As noted above, the fact that a stimu-



SINGULAR STIMULI have been shown to exist in heart pacemaker tissue by Jalife with Charles Antzelevitch of the Masonic Medical Research Laboratory. They worked with Purkinje fibers from the heart of the dog. In each panel the upper trace represents the electric potential of the fiber; the peaks indicate firing. The lower trace represents the small constant current applied to the fibers to make them fire regularly. The raised segment of the lower trace indicates the application of a small additional stimulus lasting for 200 milliseconds. If the stimulus is applied soon after the beat, the only effect is to delay the next beat slightly (*1*). If the stimulus is given a little later, the fiber stops firing but resumes its rhythm (*2*). If the stimulus is given still later, the rhythmic firing stops and is not resumed (*3*). The time when a stimulus can have such an effect is called the vulnerable phase. Stimuli with the right strength given in the vulnerable phase correspond to the black regions in the illustration on the preceding two pages. If the stimulus is given after the vulnerable phase, the next beat is advanced slightly (*4*).

lus has a consistent fundamental latency means it yields an offsetting of the usual rhythm of the pacemaker without eliminating the system's natural periodicity. The nonretraction theorem implies that there is at least one stimulus that does not yield such an offsetting. Whatever the result of applying such a stimulus is, it does not preserve the pacemaker's normal rhythm. The nonretraction theorem does not indicate what the actual result of applying such a stimulus to a physiological system might be. The subsequent beats could be unpredictable or nonexistent. The particular combination of coupling interval and stimulus strength that yields such a result represents a mathematical singularity: a "hole" in the pattern of timing.

It should be noted that the finding of the singularity depends on the existence of strong rescheduling. If the pacemaker showed only weak rescheduling, the colors on the upper edge of the rectangle would vary through one full cycle, as those on the lower edge do. In the exercise with the wheel the movement along the top would cancel that along the bottom: the wheel would make no net forward motion. As a result the nonretraction theorem could not be applied.

The topological patterns of weak and strong rescheduling apply to simple systems as well as to complex ones such as the human heart. To determine whether the predicted singularity could actually exist in a physiological system my student Eric Best employed a computer to simulate the action of a very simple system capable of periodic "beating": a single nerve cell. The simulation was based on a well-known set of differential equations formulated by Alan L. Hodgkin of the University of Cambridge and Andrew F. Huxley of University College London to describe the propagation of the action potential along the giant axon (an exceptionally large nerve fiber) of the squid. Best used the Hodgkin-Huxley equations to construct a model of rhythmic firing in the squid axon. He then simulated the delivery of a range of stimuli while varying the coupling interval. The resulting latency was recorded.

The giant axon of the squid is sensitive to both excitatory and inhibitory stimuli. Therefore the simulation yielded two rectangles of the kind described above, one rectangle for each type of stimulus. The two rectangles can be joined along the horizontal axis corresponding to a stimulus of negligible intensity. The measured latencies were marked in the two rectangles and contour lines were drawn connecting points with the same latency, much as contour lines on a topographical map connect points with the same altitude. Contour lines that represent latencies are called isochrons. When the isochrons had been drawn in the rectangles, the contours were filled in with colors indicating fractions of the natural period.

The results of the plotting and coloring were quite exciting. In both rectangles the isochrons converge on a "black hole": a singularity where the latency is not defined. A full cycle of isochrons radiates from the perimeter of each black hole, a fact that is of considerable significance in connection with fibrillation. When the two rectangles are joined along the horizontal axis, the isochrons stream in an orderly way from one singularity to the other. Although the nonretraction theorem guarantees only the existence of a singular point, the black holes in the simulation were substantial in area. The reason is that the conditions of the simulation were adjusted to overcome the difficulty of finding a singular point. As we shall see, the conditions in the heart also lead to the expansion of the singularity into a substantial region.

Best's findings have since been confirmed in actual biological systems by other workers. John M. Rinzel of the National Institute of Arthritis, Metabolism, and Digestive Diseases and Rita Guttman of the Woods Hole Marine Biological Laboratory found both excitatory and inhibitory singularities in the squid giant axon. Jalife and Charles Antzelevitch of the Masonic Medical Research Laboratory found excitatory and inhibitory singularities that terminate the spontaneous firing of the Purkinje fibers of the dog and the sinus node of the cat.

It appears that the periodicity of any pacemaker capable of both weak and strong rescheduling can be eliminated

by a brief stimulus with the right magnitude and timing. The contraction of the human heart, however, is spatially distributed, and its arrhythmias, or departures from the usual rhythm, often have the form of rapidly circulating waves rather than unpredictable timing or failure to fire. In particular, fibrillation is a spatial disorganization of the normal contraction. Can the discovery of topologically singular points, fascinating as it is, throw any light on the cause of fibrillation?

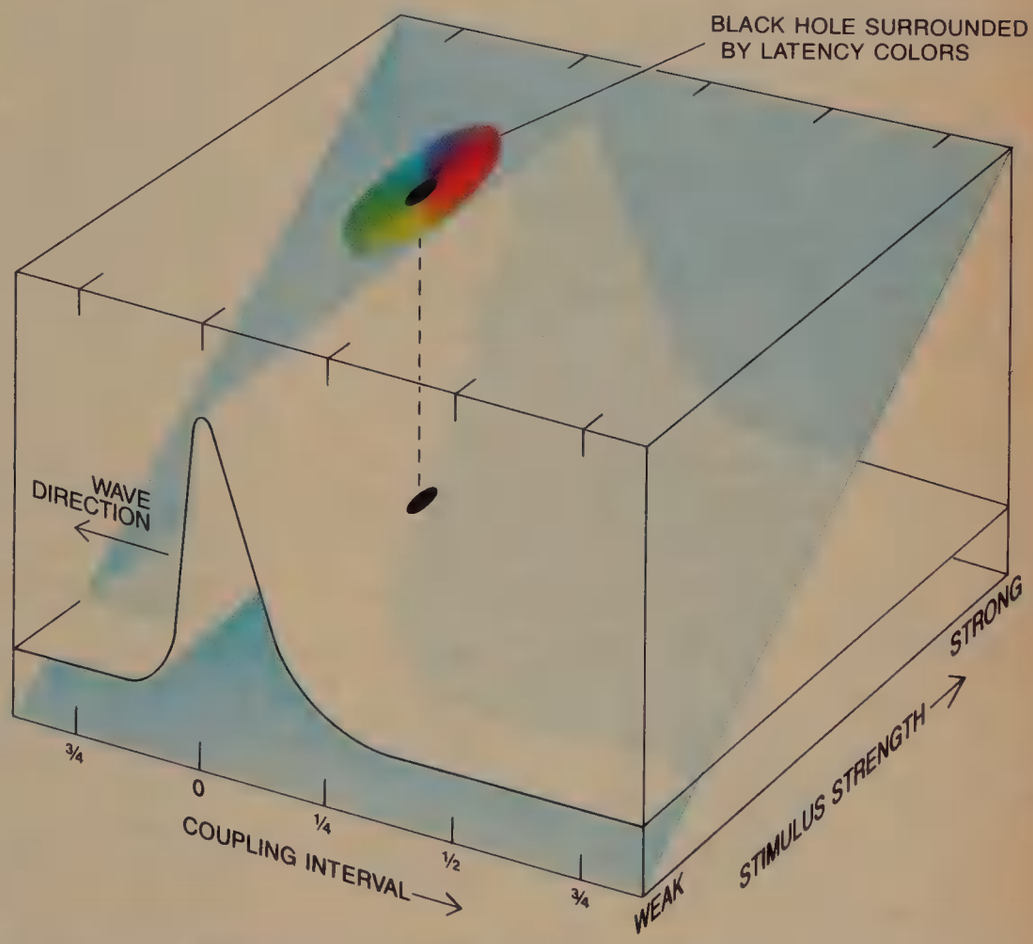
There is a connection between the singularities and circulating waves, but in order to understand the connection it is necessary to consider how waves of electrical impulses propagate in the heart. In heart muscle the electrochemical impulses that trigger a contraction propagate without attenuation. Each fiber restores the passing signal to its full strength before the signal is transmitted to a neighboring cell. Therefore if the signal could be deflected into a circular path, it could continue propagating around the heart indefinitely. The phenomenon is referred to in modern clinical terminology as reentry.

The continuous circulation of an electrical wave in the heart can become established only if the circular path of the wave is too long to be traversed within the refractory period that follows each action potential. If the period of the circulating wave were shorter than the refractory period, the wave would disappear after a single circuit because it would then encounter fibers still unable to fire. It is now known that waves can circulate on an appropriate path in the heart. Indeed, waves circulating around an obstruction such as the opening of a blood vessel or a patch of dead fibers have been shown to underlie several forms of arrhythmia.

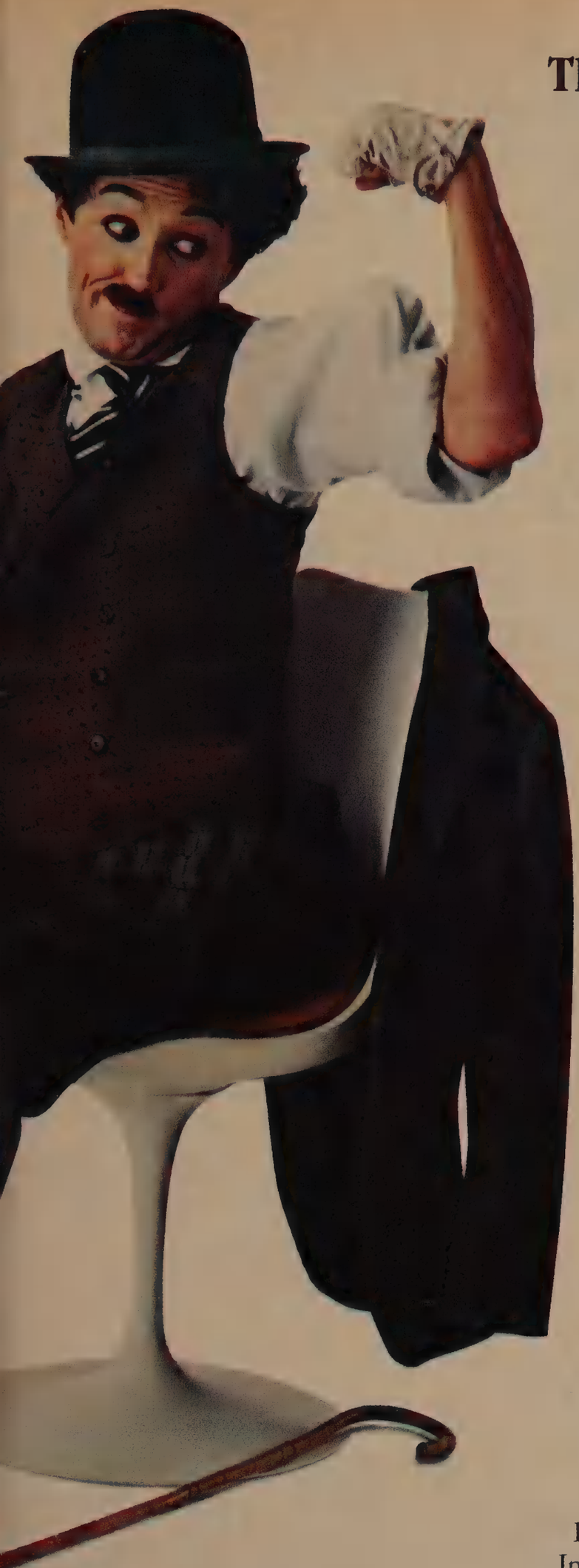
It was long thought such waves could not circulate in unperforated tissue because they would quickly cancel themselves by propagating across the tissue in the center of the circular path. Recent work in which computers have been utilized to simulate the action of heart tissue has shown, however, that a wave can circulate even in continuous tissue such as healthy ventricular muscle. Gordon K. Moe of the Masonic Medical Research Laboratory, J. A. Abildskov of the University of Utah Medical Center at Salt Lake City and Werner C. Rheinboldt of the University of Maryland at College Park were the first to simulate an excitable medium like heart muscle and show that circulating, or rotating, waves can exist. V. I. Krinskii of the Institute of Biophysics of the U.S.S.R. Academy of Sciences and his colleagues have found that a single rotating wave, or rotor, can multiply, fragment and spread if the tissue has inhomogeneous areas such as those that frequently result from obstructions of the flow of blood to the cardiac muscle. Such rotors can circulate faster and in a much smaller space than had been thought possible. This work yields the best theoretical model of fibrillation.

Allessie and his colleagues Felix I. M. Bonke and F. J. G. Schopman were the first to observe rotors in actual heart tissue. The Dutch workers were able to induce a wave circulating about 10 times per second through an unperforated piece of rabbit tissue. They employed an array of moving electrodes to record the impulses in the tissue. The position of the wave front was plotted as a sequence of isochronal contours. As the wave rotates, the inner endpoint of each isochron describes an irregular loop less than a centimeter across. Similar isochronal maps of ventricular fibrillation have been obtained by M. J. Janse and F. J. van Capelle of the Hospital of the University of Amsterdam.

As I have mentioned, the nonretraction theorem does not predict what kind of arrhythmia might be caused by a stimulus in the singular region. The theorem predicts only that such a stimulus exists. The topological argument is therefore quite limited in predictive power. On the other hand, the topological principles have a more general interpretation than I have given them so far. The principles of pacemaker timing given above could apply to a spatially distributed system as well as to a single contracting mass. In a spatially extended system the isochronal contours in the rectangular graphs could be mapped onto a region of tissue. They would represent the electrical activity distributed over the tissue at a given moment rather than the accumulated results of stimu-



AUTHOR'S HYPOTHESIS of how fibrillation could begin includes two gradients: one of stimulus strength and one of timing. The transverse overlapping of the gradients results in a mathematical singularity in the heart. The illustration shows a small rectangular area of cardiac tissue in schematic form. The wave passing through the tissue from right to left is an impulse from the heart's pacemaker. As the wave passes, the fibers fire in turn. The fibers to the left of the peak fired almost a full period ago and will soon fire again. Those at the peak are just firing. The fibers immediately to the right of the peak have just fired. Thus all values of the coupling interval from zero to one period are laid out along the front of the rectangle. One of the nerves that modulate the action of the heart has endings that are densely concentrated near the rear of the rectangle. The endings become sparser with the distance toward the front of the rectangle. As a result a stimulus from the nerve will be strongest at the rear of the rectangle; its strength will decrease with the distance toward the front of the rectangle (*tilted plane*). Thus the latency diagram, with coordinates of stimulus strength and coupling interval, is laid out on the rectangle of tissue. It follows that somewhere in the rectangle there is a combination of stimulus strength and coupling interval that does not yield a consistent latency. Radiating from the singularity is the complete sequence of isochrons. When a stimulus from the nerve strikes the rectangle, the fibers around the singularity begin to fire in the order of their latencies, which corresponds to the sequence of isochrons. This is an ideal setup for starting a rotating wave.



The IBM Personal Computer XT. More power to the person.

Plenty of muscle. That's what the new IBM Personal Computer XT means to a person with heavyweight data to manage. Because one of the XT's many strong points is a 10-million-character fixed disk drive that helps give you the power to pump more productivity into your business. What's so special about a fixed disk? Exactly that. It's *already* fixed inside the system, with the capacity to store the facts, figures, names and numbers you need to work with. (Rather than go from diskette to diskette, store up to 5,000 pages of text or up to 100,000 names and addresses in one place.) Yet there's more built into the XT than its fixed disk. Reliability and quality are built in as well. Plus more than 30 years of IBM experience. A new level of price/performance. And a remarkable compatibility of both software *and* hardware with the original IBM Personal Computer. So, with the introduction of XT comes a special tool designed to help you be more productive in high-volume applications.

WHAT'S THE DIFFERENCE?

IBM Personal Computer	IBM Personal Computer XT
BASE SYSTEM*	**BASE SYSTEM***
User Memory 64KB (expandable to 640KB)	**User Memory** 128KB (expandable to 640KB)
Auxiliary Memory Up to two 5¼" 160KB/180KB or 320KB/360KB diskette drives optional	**Auxiliary Memory** One 10 megabyte fixed disk drive and one 5¼" 360KB diskette drive standard

*An expansion unit can also be added to both 16-bit (8088) systems for 6 more system expansion slots. Added to the IBM Personal Computer, it can house two 10 megabyte fixed disk drives. Added to the IBM Personal Computer XT, it can house one additional fixed disk drive for a total of 20 megabytes.

Another tool for modern times to keep you going strong.
To find out where you can see the IBM Personal Computers, call 800-447-4700.
In Alaska or Hawaii, 800-447-0890.

IBM®

lating a single fiber or a single pacemaker in many different trials over a period of time. The logic of the topological argument would be unchanged by such a transformation. Each point in the graph would still correspond to a combination of stimulus strength and coupling interval, but the point would also represent a position in the system. The color of the point would give the latency at that position. In addition suppose the spatially distributed rescheduling action of the human heart turns out to resemble the rescheduling of the squid axon as described in the Hodgkin-Huxley equations and simulated by Best. The crucial part of the resemblance would be the existence around each singularity of the full cycle of isochrons: the "rainbow" sequence of colors that surrounds the black hole.

How could such a rainbow be formed on the surface of the human heart and what would its consequences be? Consider a small rectangular piece of the muscular wall of the ventricle. Successive electrical waves from the pacemaker move continually across the tissue. For the sake of visualization it is convenient to imagine that the wave fronts are passing from right to left through the rectangle. Now suppose the length of the rectangle is roughly equal to the wavelength of the impulse. There will be one peak of the wave in the rectangle. The fibers at the peak are just firing. The fibers slightly to the left of the peak fired almost a full cycle ago and are about to fire again. The fibers to the right of the peak fired a short time ago as the peak passed; the fibers farther to the right fired longer ago. Thus all possible coupling intervals from zero to one natural period are laid out along the front of the rectangle.

Spread throughout the tissue are endings from several major nerves. The endings, however, are not evenly distributed. The strength of a nerve impulse is proportional to the density of the endings. Suppose the region at the back of the rectangle has a higher density of endings than the region at the front. An impulse from one of the nerves will therefore decrease in strength from the back to the front of the rectangle.

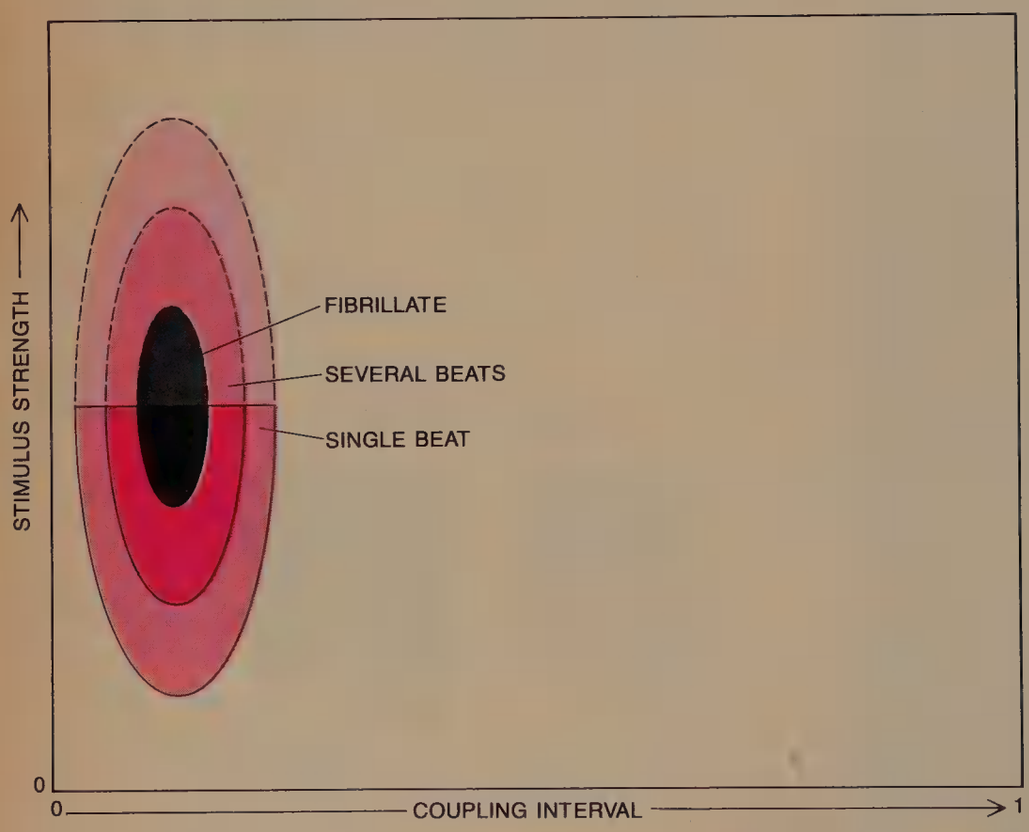
Now, when a nerve impulse enters the rectangle, the gradient of its strength will be transverse to the gradient of coupling intervals along the path of the pacemaking wave. The result of such an arrangement is that the rectangular graph of coupling interval and stimulus strength is laid out in an orderly way on the rectangle of heart muscle. If these conditions are met, then somewhere in the block there must be a black hole, a point where the coupling interval and the impinging stimulus do not give rise to any well-defined latency.

If, as I have hypothesized, the rescheduling action of the heart resembles in its essentials that of the squid axon, then arrayed around the black hole there is a rainbow of isochron colors standing for a complete sequence of latencies. Each isochron runs through cardiac fibers that fire after the same delay following the stimulus from the nerve. Thus the stimulus could yield a rotating sequence of firings around the point that represents the singularity.

As visualized here, the rainbow of isochron colors would be confined to about a square centimeter, which is roughly the area occupied by the rotors observed in heart tissue. It is difficult to imagine a better setup for starting a rotating wave than such a black hole surrounded by the rainbow of latencies. Furthermore, if the cardiac tissue includes many inhomogeneities, the rotor would be shattered into many small waves. By this process the rotor generated by a rescheduling singularity could ultimately be transformed into the catastrophic arrhythmia of fibrillation.

The gradients of timing and stimulus strength that exist in profusion in the human heart could make it easier for fibrillation to start. To understand how this could be so, consider the following argument. The topological argument guarantees only the existence of one singular point: a unique combination of coupling interval and stimulus strength. If the human heart were made up of an ideal tissue in which all impulses were transmitted instantaneously and all fibers contracted simultaneously, then it would be necessary to duplicate that combination exactly to cause fibrillation. The black hole would be an infinitesimally small point.

The heart of a large mammal such as man, however, is very different from such an idealized organ. As noted above, the timing of firings and the strength of stimuli vary considerably even across short distances. As a result many combinations of stimulus strength and timing can give rise to a rotor. For example, if a particular stimulus is too strong to generate a rotor at the point where it is applied, some distance away from that point it may have the right strength. Similarly, if the timing of the stimulus is not correct at the point of application, it may well be right a short distance away.



TARGET shows the stimuli that can cause fibrillation in intact hearts. The upper part of the graph is based on qualitative clinical data; the lower part is based both on such data and on more precise information from work with animals. In the outer ring are stimuli that cause an extra beat when they are applied to the heart. In the middle ring are stimuli that cause several extra beats. In the inner area are the stimuli that cause fibrillation. The topological argument presented here guarantees the existence of only one combination of stimulus strength and coupling interval that could start a rotating wave at a specific point on the heart. In the actual heart, however, at any one moment there are gradients of timing and stimulus strength. Such gradients increase the range of stimuli that can give rise to a rotating wave somewhere on the heart. If the applied stimulus is too large to give rise to a rotating wave, it could have the right strength a short distance away. If its timing is not right at the point of application, it could be right a short distance away. The many combinations of stimulus strength and timing that could give rise to a rotating wave in an intact human heart are shown in the bull's-eye of the target.

Imagine again a graph with the coupling interval on the horizontal axis and the stimulus strength on the vertical axis. On the graph are plotted the stimuli that can cause fibrillation. For the heart composed of an ideally uniform tissue the target created in this way is a single point. For the actual heart the target is enlarged considerably and is therefore much easier to hit. Work with animals shows there is indeed a substantial region of the graph where fibrillation can result; stimuli outside the target region do not cause fibrillation. The principle is exemplified in cases of accidental death by electrocution. It has also been employed to stop the beating of the heart during delicate open-heart operations.

In his final paper George Mines made three generalizations about fibrillation that have since been borne out by much clinical experience. First, fibrillation can be triggered by an untimely electrical impulse, that is, one coming in the vulnerable phase. Second, fibrillation frequently involves the reentry of circulating impulses. Third, fibrillation is favored by almost any departure from spatial uniformity in the target tissue or in the stimuli from the nerves that infiltrate the tissue. The topological argument given here elucidates the connections among these three empirical generalizations.

For the moment this is as far as either clinical observation or topology can take us. If the topological mechanism proposed here is correct, its clinical implications could be somewhat bleak. The heart is continually bombarded by electrical impulses from many sources. The possibility of a singularity resulting from such impulses cannot be eliminated without eradicating the strong rescheduling property of the pacemaker's membranes. If this is not possible or desirable, then the challenge to preventive medicine is to make the black holes as small and as inaccessible as possible.

Some of the unanswered questions suggested by the topological analysis may well lead to the most interesting clinical results. It appears that all human hearts are subject to bombardment by electrical impulses. Only some hearts, however, develop fibrillation. Under what conditions does a rescheduling singularity turn into a rotor? When is a rotor transformed into many scattered smaller waves? When does it simply die out into harmless synchronized activity? Can the conditions necessary to transform the singularity into fibrillation be identified and possibly controlled by drugs that alter the electrical properties of the cardiac tissue? These are questions that have long concerned cardiologists. By calling attention to them in a mathematical context topology may ultimately make a significant contribution to overcoming sudden cardiac death.

Better Than Jogging, Swimming or Cycling

NordicTrack

Jarless Total Body Cardiovascular Exerciser
Duplicates X-C Skiing For The Best Way To Fitness

NordicTrack duplicates the smooth rhythmic total body motion of XC Skiing. Recognized by health authorities as the most effective fitness building exercise available. Uniformly exercises more muscles than jogging, swimming, cycling and rowing.

Does Not cause joint or back problems as in jogging. Highly effective for weight control and muscle toning.

Easily Adjustable for arm resistance, leg resistance and body height. Smooth, quiet action. Folds compactly to require only 15 by 17 inches of storage area. Lifetime quality.

Used In thousands of homes and many major health clubs, universities, and corporate fitness centers.

Call or Write for FREE BROCHURE
Toll Free 1-800-328-5888 MN 612-448-6987
PSI 124 F Columbia Crt., Chaska, Minn. 55318

Which of these languages would you like to speak?

Mark the one you want to speak in 2 or 3 months' time

☐ Afrikaans	☐ French	☐ Japanese
☐ American	☐ German	☐ Norwegian
☐ Arabic	☐ Greek	☐ Polish
☐ Chinese	☐ Hebrew	☐ Portuguese
☐ Czech	☐ Hindi	☐ Russian
☐ Danish	☐ Indonesian	☐ Spanish
☐ Dutch	☐ Irish	☐ Swedish
☐ Finnish	☐ Italian	☐ Welsh

A Linguaphone Course makes you feel at home in almost any country. You'll never miss the meaning of conversations or be at a loss for words.

• You learn fast because you LISTEN to real conversations on cassettes...UNDERSTAND what you hear by following illustrated texts...HOLD CONVERSATIONS with native-born speakers. • Learn at your own pace, on your own schedule • Quickly build a good working vocabulary, develop an authentic accent. • PROVEN SUCCESS-- over 4 million Linguaphone students in 82 countries speak a second language fluently. • SATISFACTION GUARANTEED

The Language Masters

World Language Courses, Dept. 310, 313 Nolana, McAllen, Tx. 78501. (512) 687-6761

Rush me FREE INFORMATION about learning languages I have checked.

Name _____

Street _____

City _____ State _____ Zip _____

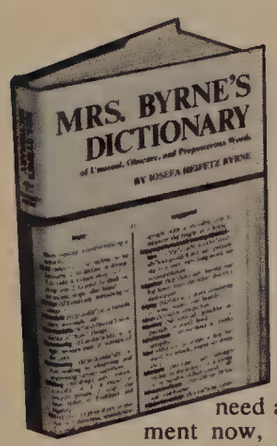
Linguaphone

For years I suffered terribly from

LETHOLOGICA

until a nice doctor friend prescribed

MRS. BYRNE'S DICTIONARY
of Unusual, Obscure, and Preposterous Words



Yes, for years I couldn't remember the right words. People I considered complete idiots were finishing my sentences for me. I found how bleak the future was when I tried a little cephalonamancy (fortunetelling by boiling an ass head), and tyromancy (fortunetelling by watching cheese coagulate.)

I went from doctor to doctor looking for help, finally becoming a confirmed iatrapistiac (one having little faith in doctors), especially when one suggested I needed a hepaticocholangiocholecystenterostomy (look it up).

Then I found *Mrs. Byrne's Dictionary*. Now I can be unusual, obscure and preposterous by turns. Now I don't need an unabridged to go along with my collegiate. For entertainment now, I browse instead of groak (watching people silently while they eat, hoping they'll ask you to join them).

Leslie Hanscom in *Newsday:* "You can dip in anywhere and come up with pay dirt . . . hundreds of words to stand your hair on end. Only a clinchpoop could scan these pages without a feeling of awe at the undiscovered boundaries of the English tongue!"

Donald B. Thackrey of *United Press International:* "A dictionary you can browse through and read like a book."

Sydney J. Harris in the *Chicago Daily News:* "There is no comparable work; I unreservedly recommend it."

At your favorite bookshop or order directly from the publisher, using the handy coupon below.

University Books Inc., Dept. SA-10 **120 Enterprise Avenue • Secaucus, New Jersey 07094**

Please send me by return mail a copy of MRS. BYRNE'S DICTIONARY for which I enclose $12.50 plus $1.50 for postage and shipping.

My name: _____

Address: _____

City: _____ State: _____ Zip: _____

THE AMATEUR SCIENTIST

What causes the "tears" that form on the inside of a glass of wine?

by Jearl Walker

When wine is left standing in a glass for a few minutes, a curious array of drops is likely to appear on the inside of the glass above the surface of the liquid, particularly if the wine has a fairly high alcoholic content. The phenomenon is known as "tears of strong wine." It was examined as early as 1855 by James Thomson, a British engineer and physicist, who concluded that it did not arise from the condensation of water on a cool surface, since the drops appeared even when the glass was at room temperature. He thought the drops formed as a result of variations in the surface tension of the wine as alcohol evaporated from a thin film of wine on the wall of the glass.

The formation of the tears has since figured in the general study of circulation patterns in liquids varying in surface tension. Although such motions have been studied for more than a century, the physical events are still not understood in detail. The motions are interesting to a physicist because they serve as a tool in the study of surface tension and the stability of circulation patterns in liquids. They are also important because they can be found in many industrial and biological processes.

Although Thomson's explanation for the tears of strong wine was basically correct, his work was apparently ignored until 1869, when the entire subject of motion resulting from surface tension was reviewed by Gustav L. van der Mensbrugghe in a Belgian journal. The priority of the research was challenged two years later in Italy by Carlo Marangoni. Thomson's early work was virtually forgotten, and Marangoni's name is now attached to liquid motions driven by surface tension.

Thomson's explanation was a simple one based on the observation that when alcohol is added to water, the surface tension of the water decreases. Hence wine has less surface tension than pure water. If the wine is exposed to air, the alcohol in it continuously evaporates, creating areas of higher tension on the surface. Such an area pulls on the adjoining liquid and starts it in motion.

The more visible formation of tears on the wall of the wineglass is due to the same kind of variation in surface tension. The surface of the liquid is curved at the wall because the surface tension at the interface of the wine and the glass pulls the wine up the wall a little way.

Alcohol evaporates from this film, increasing the surface tension of the liquid there and causing more wine to be pulled up. Thomson noticed that the additional liquid tends to form a thick ring at the top of the film. As the alcohol evaporates and the surface tension in the ring increases, the liquid begins to contract into small drops. More liquid is pulled up. More alcohol evaporates. The drops get larger. Finally a drop gets heavy enough to slide back down into the wine. Soon another drop forms in its place because evaporation continues in the film and fresh liquid is constantly pulled up. The process continues until so much alcohol has been removed that the variation in surface tension gets too small to keep the cycle going.

Thomson tested the role of evaporation by corking a partly filled vial of wine. After shaking the mixture he examined the film on the interior wall. No tears of wine formed and no motion was apparent in the film except normal drainage. Then he removed the cork, so that fresh air entered the vial. Under these conditions, he reported, "a liquid film is instantly to be seen creeping up the interior of the vial with thick or viscid-looking pendent streams descending from it like a fringe from a curtain." When the cork was in place, shaking so saturated the air in the bottle with alcohol that no more evaporation was possible. Fresh air restored both evaporation and the cycle whereby liquid was driven into motion by the variation in surface tension.

I demonstrated the formation of tears with several kinds of alcoholic beverage. First I poured a small amount of 94.6-proof gin into a watch glass. (The concentration of alcohol by volume is half the proof number.) Four tears formed almost immediately near the rim of the glass. The region between a tear and the rest of the gin was visibly wet and occasionally showed motion. I doubt that I saw gin flowing up the sides of the watch glass, since gin is transparent; it is more likely that I was seeing dust motes caught up in the flow. To improve the visibility I dusted the surface of the gin with lycopodium powder. The microscopic spores of the powder rest on top of the gin instead of sinking. Talc and other household powders that are not readily wetted by water and alcohol can serve in place of lycopodium.

With the powder in place as a tracer I could follow much of the flow of gin up the side of the glass. In general the motion was irregular, but below a spot where a tear formed the flow was mainly upward. (The irregular flow elsewhere results because the film is evaporating and thus varying in surface tension from place to place.)

A tear would eventually slide abruptly down into the gin or would descend gradually until it touched it. In the latter case the drop jerked a little way back up the wall after touching the gin. The reason is mainly that the drop loses liquid when it touches the surface. It is then lighter and more responsive to the surface tension pulling upward than to its weight pulling downward. The jerk is also promoted by the sudden difference in surface tension at the interface of the drop and the gin.

A drop touching the gin injected a small jet of liquid into the gin. The jet is visible because it has an index of refraction different from that of the rest of the gin and therefore bends the light rays crossing through it. The difference in refractive index probably results from two factors: the jet has both more alcohol and greater density than the gin around it. The jet is denser because the liquid has been cooled by the evaporation that causes tears to form.

I next poured about a cup of 80-proof rum into a porcelain bowl and added a small amount of chocolate flavoring at one spot near the edge of the liquid. Beautiful purple patterns raced through the rum. The colored lines were quite lively, apparently driven by strong variations in the surface tension at the top of the mixture and by circulation systems within it.

I put the bowl in a warm oven (its temperature somewhat below 93 degrees Celsius, or 200 degrees Fahrenheit). Tears soon formed on the inside surface of the bowl. With a flashlight and a small magnifying glass I examined the edge of the rum. Small, undissolved particles left from the flavoring served to trace the liquid motion. The particles darted toward and away from the edge, revealing a vigorous and complex circulation of liquid there.

Food coloring turned out to be a bet-

Movements caused in vodka by surface tension, as traced by red wine

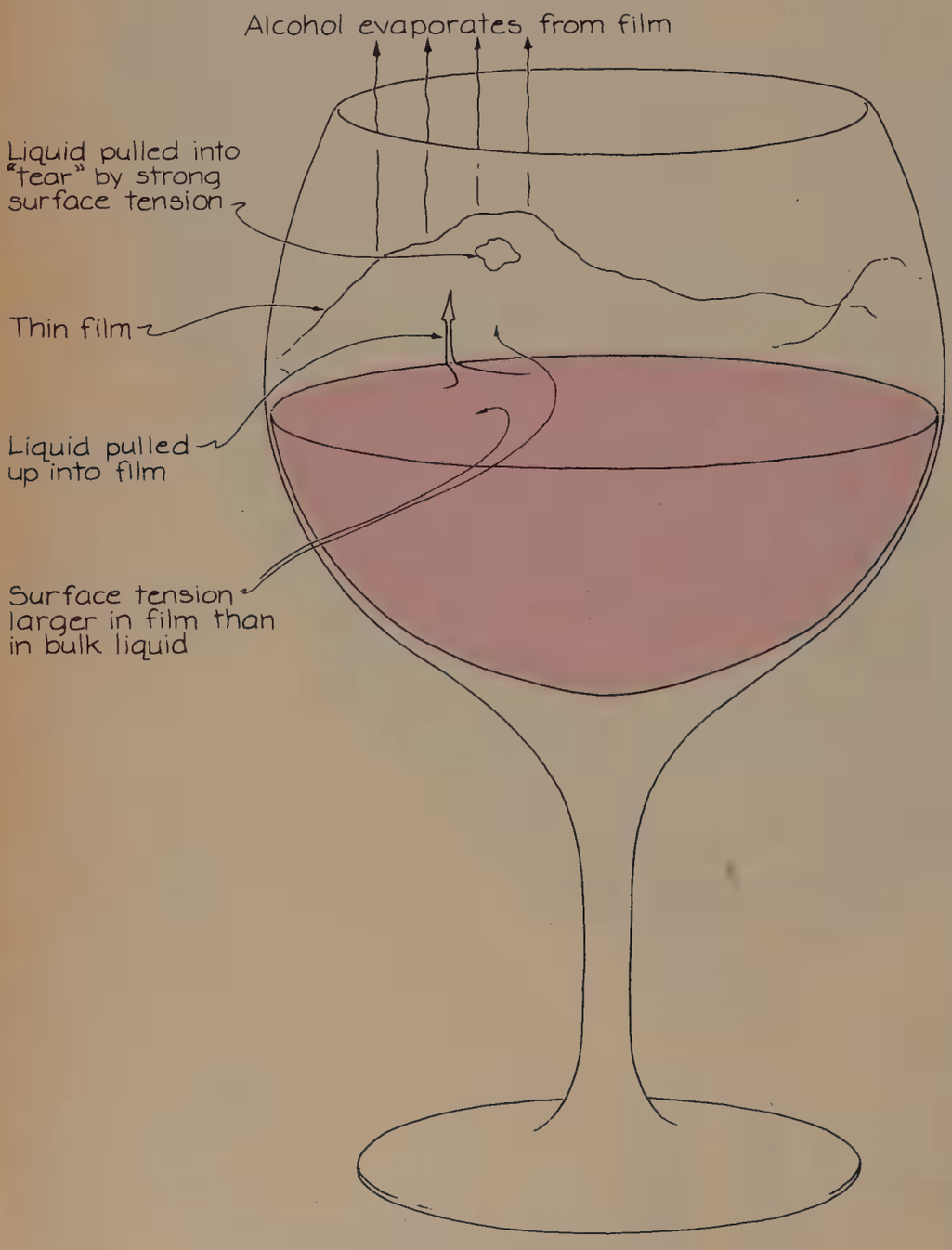
ter tracer. A small drop added to isopropyl alcohol in a porcelain cup colored one section of the perimeter. I watched as a small colored section of the alcohol reached the perimeter of the cup and began climbing through the invisibly thin film of alcohol and water lining the wall, finally entering a tear that had already formed on the wall.

I next made a simple but crucial test of Thomson's hypothesis that the drops appear only because of a difference in the surface tension in the film on the wall. Suppose the alcohol contains no water. Then the surface tension would hardly vary as the alcohol evaporated on or near the wall. Indeed, it would not vary at all except for the changes caused by the fact that the liquid cools as it evaporates.

Working with a pair of identical cups, I poured a quantity of 70 percent isopropyl alcohol into one cup. In the other I poured half as much alcohol and then added enough water to bring the surface level up to that of the first cup. I gave both cups a brief swirl to wet the sides. Many tears formed quickly above the diluted sample. Only a few small tears appeared above the concentrated sample. Thomson was right: water is necessary for a variation in surface tension.

Any factor that promotes evaporation also aids the formation of tears. A warm environment, direct sunlight and a wide, shallow container help. I made a large-scale demonstration of tear formation employing a serving platter with sides that curved upward. A tall glass with a small amount of liquid works poorly because the evaporation is slow, unless you first wet the entire inside wall. Then when the alcohol is poured into the glass, the liquid film rapidly climbs the wall, rising about as high as the remaining wetness. The appearance of tears several centimeters above the alcohol mixture seems almost magical. A watcher might say the tears are actually drops of condensation. To prove their origin you can color the liquid at the bottom of the glass. The tears will be colored too.

You might like to investigate the formation of tears with various alcoholic drinks. Beers made in the U.S. do not work, at least in my experiments, apparently because they contain too little alcohol. Can you find liquids other than alcohol and water that form tears?



The forces that create "tears of strong wine"

As I was investigating the tears of strong wine I was writing the article about Middle Eastern coffee that appeared in this column last month. The drink consists of water, sugar and finely ground coffee grains, all brewed together and served in a cup along with the grains. I prepared so much of the stuff that I began to pour it into bowls instead of drinking it. When I was cleaning up one morning, I discovered a pattern in a bowl of coffee left from the previous night. Along the edge of the liquid, just below the surface, lay a neat array of dark lanes of coffee silt. They and the intervening clear lanes were each a few millimeters wide.

I was stumped by the pattern. Not one of the hundreds of unfinished cups of coffee I had left lying around had ever shown it. Is Middle Eastern coffee somehow special? I prepared another batch and left it in the same kind of porcelain bowl. I saw nothing interesting in the period of rapid evaporation from the hot surface, but in rechecking over the next several hours I began to notice the same pattern developing. Although I sat patiently with a flashlight and a magnifying glass, I could not actually see the lanes growing. Since I had tears of strong wine on my mind, I wondered if the patterns could be related. No, the liquid showed little movement up the side of the bowl.

Perhaps this pattern has been reported before, but I can find no mention of it. A silt such as the residue in Middle Eastern coffee is required, as is sugar. I prepared two bowls of the coffee, one bowl with sugar and one without. The next morning the one with sugar had a fine pattern and the other did not.

Evaporation is also needed. I put a bowl of the coffee in a warm oven and quickly got the pattern. Erasing it with a gentle swirl, I covered the bowl with plastic food wrap. After a while I shone a flashlight through the plastic. The pattern was there. I gave the bowl another swirl and waited again. The pattern would not re-form.

Apparently the pattern developed the first time because the water was able to evaporate into the air above the coffee. Thereafter the saturation of the air with water vapor prevented further evaporation. The high concentration of water vapor trapped on the inside of the plastic wrap was demonstrated when I opened the oven door for a look; the cool air from the kitchen caused rapid

72 GOLF

TELECOMMUNICATIONS

"72 Golf" is the Army's name for a soldier who has learned to operate sophisticated data telecommunications equipment. From handling computers to handling computer-driven tanks, you can choose from over 300 career training areas. We'll guarantee you the training you qualify for. Which means you get to serve your country in the way that best suits your talents.

So if it's not 72 Golf, it might be 93 Hotel.*

ARMY.
BE ALL YOU CAN BE.

Air Traffic Control

For more information, call toll free, 800-421-4422. In California, 800-252-0011. Alaska & Hawaii, 800-423-2244.

A carat or more.
Because you were never very good at fractions.

Your jeweler can show you many diamond solitaires of a carat or more. This necklace features a 1.02 carat diamond.

A carat or more—one in a million.

Every diamond is rare.
But of all diamonds found, a solitaire of a carat or more is only one in a million.

And, like love, becomes more precious with time.

A miracle among miracles.
Born from the earth. Reborn on a woman.

The extraordinary diamond of a carat or more.

Show the world you couldn't have made it without her.

A diamond is forever. De Beers

condensation on the inside of the plastic.

I then set about investigating the formation of the pattern by preparing many samples of Middle Eastern coffee and pouring them into a wide beaker and an assortment of watch glasses. The vertical wall of the beaker enabled me to examine the sedimentation of the grains in the coffee with the aid of a strong flashlight. I saw nothing interesting in the beaker and expected nothing but the usual pattern in the watch glasses. Surprisingly, every batch in the watch glasses failed to form a pattern. I substituted plastic lids shaped like the watch glasses but still could not produce the patterns. I was stymied.

Back with the porcelain bowls, I began to search for circulation patterns near the edge of the liquid by adding some kind of tracer. In one trial I injected a small amount of food coloring by inserting the tip of a hypodermic syringe just below the coffee level. In another trial I dusted the top surface with lycopodium powder. I could not discern any circulation, but when I left the bowl at room temperature overnight, the pattern appeared.

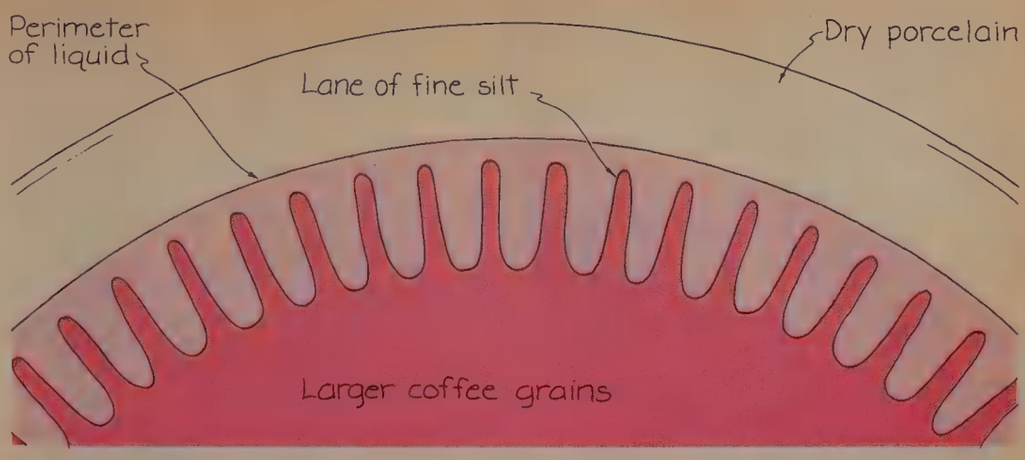
I tried for similar patterns with other mixtures. Neither tea leaves nor grains of sand produced a pattern in sugar water. Apparently the pattern requires fine silt lying on the bottom; tea leaves and sand are too big.

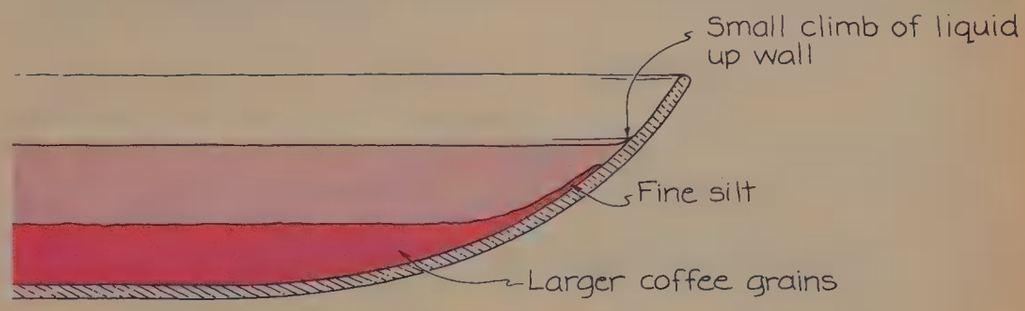
With a normal mixture of Middle Eastern coffee in a bowl I scraped the silt at one side of the bowl and then waited for the pattern to form. It did form but not in the area I had cleared. This result gave me a useful clue to the nature of the pattern. Apparently the silt had to be near the edge of the liquid. Whatever causes the pattern does not appreciably transport the silt up the wall of the bowl and toward the liquid's edge.

I made a mixture of Middle Eastern coffee and gin with a small amount of red food coloring as a tracer. The alcohol was poured in slowly after the coffee had cooled for a while. Soon tears formed on the upper reaches of the wall but the coffee pattern failed to appear.

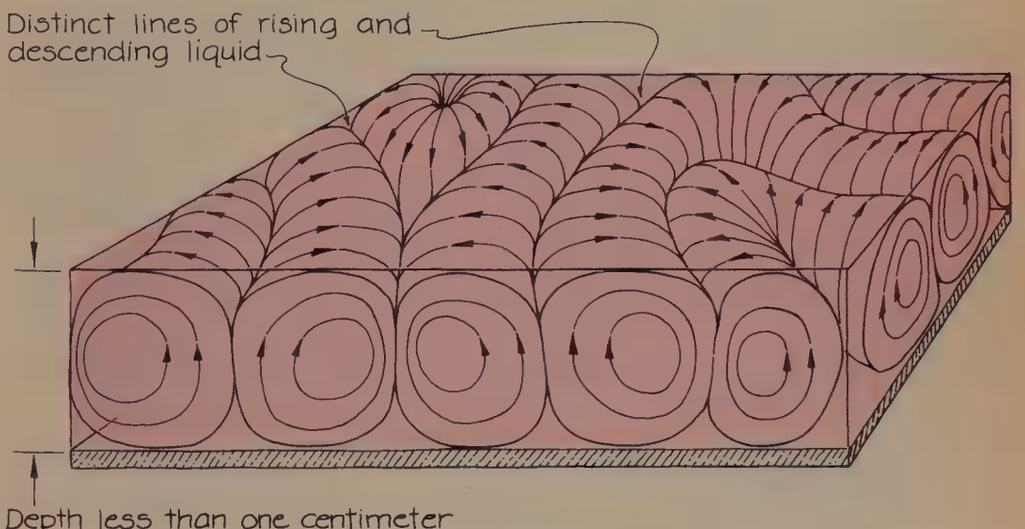
From this test, however, I received my next clue to the cause of the pattern. Whenever a tear slid down the wall and into the drink, it cleared a lane through the sediment just below the spot where it entered the liquid. In this way the drops soon produced a pattern resembling the patterns I had seen only with the coffee. The spacing in the pattern was about 10 times too large, and the pattern varied a lot because the points where the tears entered the liquid shifted.

Perhaps a circulation system at the shallow edge of normal Middle Eastern coffee generates the pattern in the silt. Two mechanisms can drive such a system in a naturally evaporating liquid. (Natural evaporation takes place without an additional source of heat, such as

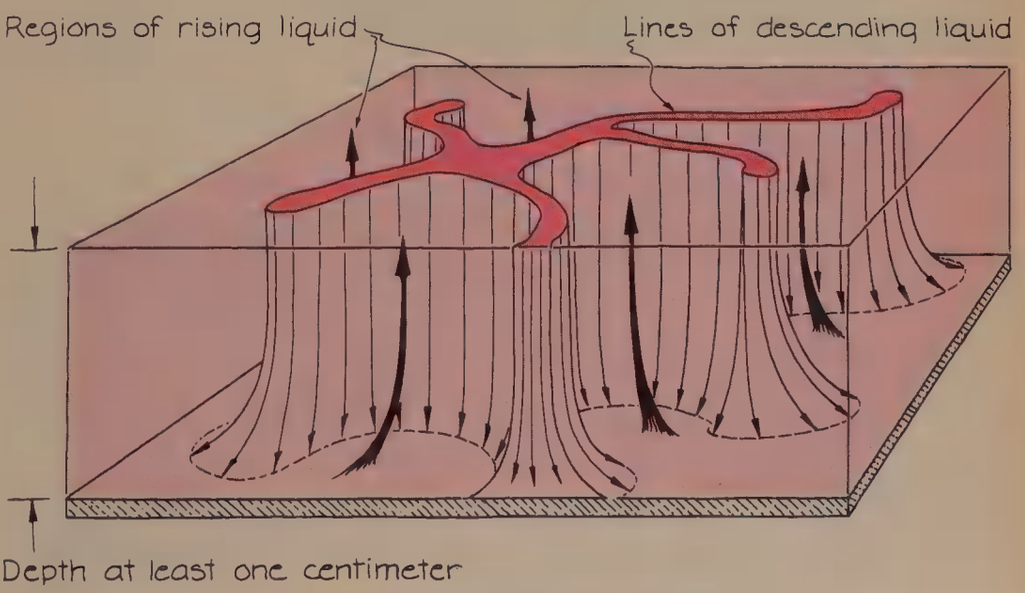




A serrated pattern on the edge of silt from Middle Eastern coffee



Circulation in a shallow liquid



Circulation in an evaporating liquid

a stove, to supply the energy for vaporization.) The liquid might move because of the variation in surface tension over its surface. It might also move because the liquid in an upper level becomes denser than the liquid under it. A circulation system is created when the conditions altering surface tension and density recur as fresh liquid is brought in.

For example, circulation cells can be seen in a cup of hot coffee. Hot liquid from the bottom rises to the top, where it cools by evaporation as it flows a short distance over the surface. Then it sinks in narrow, crooked lines to the bottom. The cycle continues until the coffee gets too cool to sustain it. The cells in hot coffee are randomly shaped and constantly changing. In some circumstances and with certain liquids they have ordered shapes and can be quite beautiful.

Natural evaporation in liquids drives circulation cells because the evaporative cooling increases the density and surface tension of the liquid. Either the increased surface tension pulls in fresh liquid or the denser liquid sinks and is replaced by fresh liquid. It is usually difficult to determine which of these mechanisms is primarily responsible for a circulation system.

According to research published a number of years ago by J. C. Berg and Michel Boudart of the University of California at Berkeley and Andreas Acrivos of Stanford University, neither mechanism operates unless the liquid is deep enough. For water the required depth is one centimeter. Shallower water displays no circulation systems; deeper water has the large circulation cells that can be seen in hot coffee.

Several other liquids investigated by Berg, Boudart and Acrivos had circulation systems even when they were quite shallow. The investigators monitored each liquid as the depth was increased. In all liquids except water circulation systems first appear when the depth of the liquid is about two millimeters. The systems are "two-dimensional, worm-like roll cells." As the depth is increased to one centimeter the cells increase in width and the lines of descending liquid become more distinct. At a depth of one centimeter the circulation cells resemble those in hot coffee. The regions where the liquid rises to the surface are then indistinct.

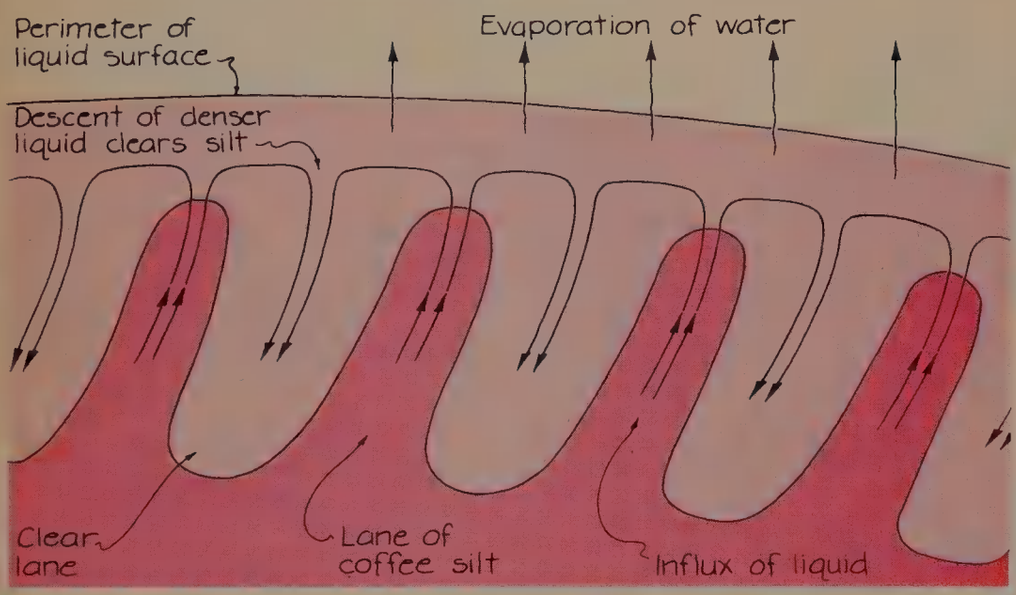
Why is water so different? Trace contaminants from the air, the container or even the experimenter spread a monolayer (a layer one molecule thick) over the surface, preventing any circulation generated by variations in surface tension. Then only circulation generated by variations in density is possible. Apparently a monolayer of contamination can stabilize water shallower than one centimeter.

This finding disappointed me. It implied there should be no circulation system near the shallow edge of evaporating Middle Eastern coffee. Yet the coffee silt revealed a pattern regular enough to have been made by hand. The coffee was certainly not pure water. It contained a lot of sugar and oil, and it was exposed to the air for hours. The top surface was undoubtedly coated with contamination at least one molecule thick. In some places I could even see tiny pools of oil.

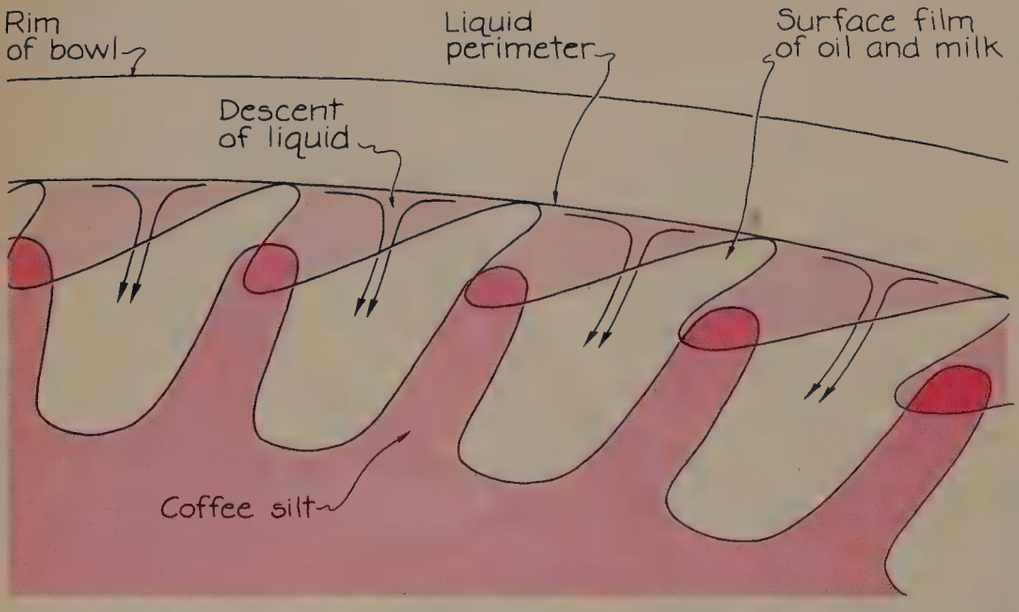
Ultimately I succeeded in putting all the clues together. The pattern requires evaporation, but because of contamination it cannot arise solely from a variation in surface tension. The pattern appears in the fine silt just below the edge of the liquid. If the silt is too far below the edge, no pattern develops. It forms only on a surface of moderate slope, not on the vertical wall of a beaker and not on the shallow slope of a watch glass.

Contamination over the surface of the coffee surely interferes with the natural evaporation of the water from the bowl but may not reduce the evaporation at the edge as much. When some of the water evaporates from an area of liquid at the edge, the remaining liquid is made denser because the concentration of sugar and oil is then higher. This section sinks. As it descends it sweeps out a lane in the silt. Other liquid flows in to replace it. Since the circulation depends on evaporation, it is normally quite slow. Part of the flow can be toward the edge along a lane of silt. If the surface is not too well stabilized by contamination, part of the flow can be along the surface.

As evaporation continues more of the clear lanes are swept clean of silt. Some of the intervening lanes of silt may be extended toward the edge by the gentle circulation system, but the diffuse flow does not carry the silt far. The pattern is enhanced by the sugar, which makes the section of liquid left by the evaporation of water near the edge denser. The pattern appears only on moderate slopes. If the slope is too steep, the circulation has no chance to sweep out lanes of silt. If



How the circulation in coffee silt causes serrations



The formation of a pattern in a surface film

the slope is too gentle, the liquid layer near the edge is too shallow for circulation. I suppose the layer of contamination is then too close to the layer of silt to allow any stabilized flow. The descent of the dense liquid from the edge may also be too gradual for the formation of a pattern.

I did one final experiment, mixing a teaspoon of milk into my usual Middle Eastern coffee. The oil in the drink and in the milk made a film over the surface. As a pattern developed in the silt along the edge, this film formed into a similar pattern. Above the silt lanes was the film, extending to the liquid edge. Above the clear lanes the surface had no film. This arrangement seems to fit exactly the circulation system I had visualized at the edge of the evaporating coffee.

Another observation may support this analysis. I had made instant coffee with milk (the drink is often called white coffee) and had not finished it. The next morning a pattern of white lines streaked the liquid surface. Another pattern of distinct radial lines appeared in the liquid left in the spoon with which I had stirred the coffee. The depth of the liquid did not seem to matter. It was about a centimeter in the cup but only a few millimeters in the spoon.

To investigate these patterns I made another preparation of white coffee and poured the mixture into several watch glasses and one metal spoon. I dusted one watch glass with lycopodium powder. Off and on for about 10 hours I checked the samples, particularly near the edge, where I employed a strong flashlight and a magnifying glass. Large patterns of white lines appeared after several hours of evaporation, but I could detect no consistent motion of the lycopodium powder. In one watch glass the pattern resembled the veins in a leaf. In another it consisted of parallel lines across the entire surface. One pattern had what looked like an upswelling because the white lines extended outward from a single point. If I gently disturbed one of these patterns with the tip of a spoon, it usually reestablished itself within an hour or so.

Near the edge of the liquid in a watch glass fine-scale patterns appeared. The hair-thin lines of dried milk were roughly perpendicular to the edge. In the spoon the shallow coffee developed the same kind of fine-scale patterns along the edge of the liquid until evaporation had reduced the depth to less than a millimeter. Sometimes the radial lines in the spoon became pronounced and more widely spaced. The coffee itself was apparently not necessary, but it did facilitate my observations. Thin layers of milk evaporating from a watch glass displayed the same kind of fine-scale patterns along the liquid edge.

I believe the two types of patterns reveal the circulation systems in the evaporating liquid. The larger patterns in coffee cups resemble the wormlike cells. As the liquid flows across the surface and then descends, milk particles collect over the lines of descent. Hardening there, they mark the paths of descent of the coffee. In addition they may help to stabilize the location of the circulation cells, otherwise the cells would be no more stable than those in hot coffee. The cells also appear to be coupled to the gentle circulation of air over the liquid surface. When a bowl of coffee is exposed to the cool air moving downward from a cold window, the same milk patterns develop.

The fine-scale patterns along the edge resemble the patterns in coffee silt and may be due to the same circulation system. Once these milk lines are established, they might stabilize the circulation along the edge. As the water continues to evaporate from the liquid, the white lines harden on the container. When white coffee has fully evaporated from a watch glass, I can hold the glass in sunlight and see a beautiful radial pattern of lines.

A camera's view of a serrated pattern in coffee silt

How a pattern in white coffee looks to the camera

BIBLIOGRAPHY

Readers interested in further explanation of the subjects covered by the articles in this issue may find the following lists of publications helpful.

METAMAGICAL THEMAS

THE PRISONER'S DILEMMA. Anatol Rapoport and Albert M. Chammah. University of Michigan Press, 1965.

TWO-PERSON GAME THEORY. Anatol Rapoport. University of Michigan Press, 1966.

THE EVOLUTION OF COOPERATION. Robert Axelrod and William D. Hamilton in *Science,* Vol. 212, No. 4489, pages 1390–1396; March 27, 1981.

SMART WEAPONS IN NAVAL WARFARE

INTRODUCTION TO SHIPBOARD WEAPONS. Carl D. Corse, Jr. Naval Institute Press, 1975.

SHAPING THE GENERAL-PURPOSE NAVY OF THE EIGHTIES: ISSUES FOR FISCAL YEARS 1981–1985. Congressional Budget Office. U.S. Government Printing Office, 1980.

THE FALKLANDS CAMPAIGN: THE LESSONS. Secretary of State for Defence. Her Majesty's Stationery Office, 1982.

VIBRATIONS OF THE ATOMIC NUCLEUS

NUCLEAR SPECTROSCOPY. Aage Bohr and Ben R. Mottelson in *Nuclear Structure,* W. A. Benjamin, Inc., 1975.

NUCLEAR VIBRATIONS. G. F. Bertsch in *Nature,* Vol. 280, No. 5724, pages 639–644; August 23, 1979.

GIANT RESONANCES IN NUCLEI. J. Speth and A. van der Woude in *Reports on Progress in Physics,* Vol. 44, Part 3, pages 719–786; 1981.

DAMPING OF NUCLEAR EXCITATIONS. G. F. Bertsch, P. F. Bortingnon and R. A. Broglia in *Reviews of Modern Physics,* Vol. 55, No. 1, pages 287–314; January, 1983.

MICROBODIES IN THE LIVING CELL

PEROXISOMES (MICROBODIES AND RELATED PARTICLES). Christian de Duve and Pierre Baudhuin in *Physiological Reviews,* Vol. 46, No. 2, pages 323–357; April, 1966.

LOCALIZATION OF NINE GLYCOLYTIC ENZYMES IN A MICROBODY-LIKE ORGANELLE IN *TRYPANOSOMA BRUCEI:* THE GLYCOSOME. Fred R. Opperdoes and Piet Borst in *FEBS Letters,* Vol. 80, No. 2, pages 360–364; August, 1977.

THE HYDROGENOSOME. Miklós Müller in *Symposia of the Society for General Microbiology,* Vol. 30, pages 127–142; 1980.

METABOLIC PATHWAYS IN PEROXISOMES AND GLYOXYSOMES. N. Edward Tolbert in *Annual Review of Biochemistry,* Vol. 50, pages 133–157; 1981.

PEROXISOMES AND GLYOXYSOMES. Edited by Helmut Kindl and Paul B. Lazarow in *Annals of the New York Academy of Sciences,* Vol. 386, pages 1–550; 1982.

MODERN PORK PRODUCTION

SWINE SCIENCE. M. E. Ensminger. Interstate Printers and Publishers, Danville, Ill., 1970.

SWINE PRODUCTION IN TEMPERATE AND TROPICAL ENVIRONMENTS. Wilson G. Pond and Jerome H. Maner. W. H. Freeman and Co., 1974.

SWINE PRODUCTION. Jake L. Krider and William E. Carroll. McGraw-Hill Book Company, 1975.

SILENT PARTNERS: FARMERS, THEIR ANIMALS AND CONSUMERS. W. G. Pond, Lacuna Books, Interlaken, N.Y., 1978.

NEW INORGANIC MATERIALS

STRONG SOLIDS. Anthony Kelly. Oxford University Press, 1973.

INORGANIC POLYMERS. N. H. Ray. Academic Press, 1978.

MAKING THE MOST OF MATERIALS: MATERIALS AND ENERGY CONSERVATION IN THE MATERIALS PROCESSING INDUSTRIES. Science Research Council, 1979.

FLEXURAL STRENGTH AND POROSITY OF CEMENTS. J. D. Birchall, A. J. Howard and K. Kendall in *Nature,* Vol. 289, No. 5792, pages 388–389; January 29, 1981.

FLEXIBLE INORGANIC FILMS AND COATINGS. D. G. H. Ballard and G. R. Rideal in *Journal of Materials Science,* Vol. 18, No. 2, pages 545–561; February, 1983.

COMPUTER-INTENSIVE METHODS IN STATISTICS

MULTIVARIATE ANALYSIS. K. V. Mardia, J. T. Kent and J. M. Bibby. Academic Press, 1979.

COMPUTERS AND THE THEORY OF STATISTICS: THINKING THE UNTHINKABLE. Bradley Efron in *SIAM Review,* Vol. 21, No. 4, pages 460–480; October, 1979.

BOOTSTRAPPING A REGRESSION EQUATION: SOME EMPIRICAL RESULTS. D. Freedman and S. Peters. Technical Report No. 10, Department of Statistics, University of California, Berkeley, 1982.

THE JACKKNIFE, THE BOOTSTRAP AND OTHER RESAMPLING PLANS. Bradley Efron. SIAM Monograph No. 38, Society for Industrial and Applied Mathematics, 1982.

THE VARIABILITY OF ACID RAINFALL. B. Eynon and P. Switzer. Technical Report No. 58, Department of Statistics, Stanford University, 1982.

THE SOCIAL INFLUENCE OF THE MOTTE-AND-BAILEY CASTLE

MANUEL D'ARCHÉOLOGIE MÉDIÉVALE. M. de Bouard. Sedes, 1975.

ENGLISH CASTLES. R. Allen Brown. Batsford, Ltd., 1976.

LE CHÂTEAU DANS LA FRANCE MÉDIÉVALE: ESSAI DE SOCIOLOGIE MONUMENTALE. Gabriel Fournier. Aubier Montaigne, 1978.

SUDDEN CARDIAC DEATH: A PROBLEM IN TOPOLOGY

SUDDEN CARDIAC DEATH: THE MAJOR CHALLENGE CONFRONTING CONTEMPORARY CARDIOLOGY. Bernard Lown in *American Journal of Cardiology,* Vol. 43, No. 2, pages 313–328; February, 1979.

PHASE RESETTING AND ANNIHILATION OF PACEMAKER ACTIVITY IN CARDIAC TISSUE. Jose Jalife and Charles Antzelevitch in *Science,* Vol. 206, No. 4419, pages 695–697; November 9, 1979.

SUDDEN CARDIAC DEATH—1978. Bernard Lown in *Circulation,* Vol. 60, No. 7, pages 1593–1599; December, 1979.

THE GEOMETRY OF BIOLOGICAL TIME. Arthur T. Winfree. Springer-Verlag, 1980.

COMPUTER SIMULATION OF ARRHYTHMIAS IN A NETWORK OF COUPLED EXCITABLE ELEMENTS. Frans J. L. van Capelle and Dirk Durrer in *Circulation Research,* Vol. 47, No. 3, pages 454–466; September, 1980.

FIBRILLATION AS A CONSEQUENCE OF PACEMAKER PHASE-RESETTING. A. T. Winfree in *Cardiac Rate and Rhythm,* edited by L. N. Bouman and H. J. Jongsma. Martinus Nijhoff, 1982.

THE AMATEUR SCIENTIST

THE MARANGONI EFFECTS. L. E. Scriven and C. V. Sternling in *Nature,* Vol. 187, No. 4733, pages 186–188; July 16, 1960.

INTERFACIAL PHENOMENA. J. T. Davies and E. K. Rideal. Academic Press, 1963.

NATURAL CONVECTION IN POOLS OF EVAPORATING LIQUIDS. J. C. Berg, M. Boudart and Andreas Acrivos in *Journal of Fluid Mechanics,* Vol. 14, pages 721–735; 1966.

JUN 7 '83

Marygrove College Library
Detroit, Michigan 48221
PLEASE DO NOT REMOVE

SCIENTIFIC AMERICAN

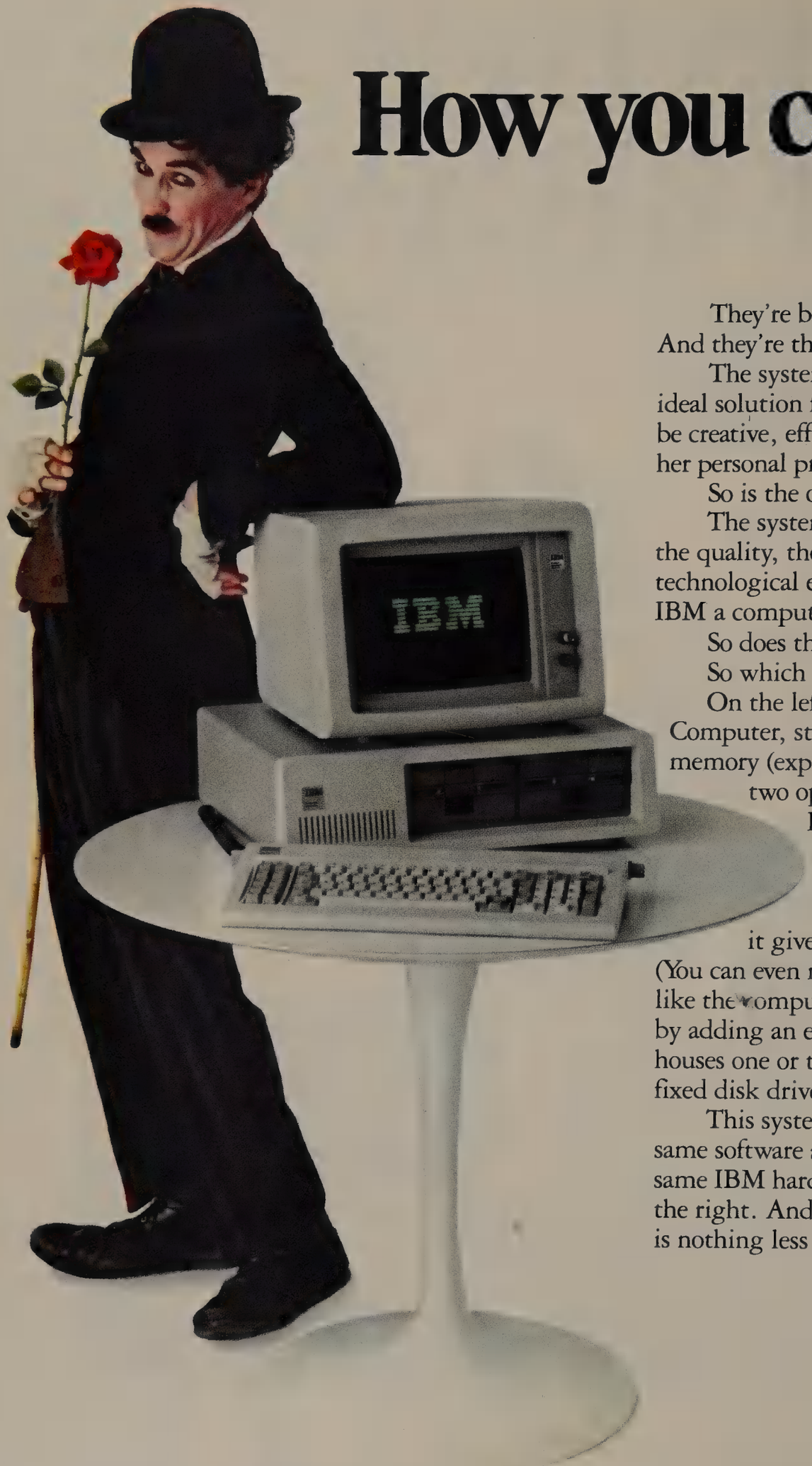
THE SLIDE FASTENER

$2.50

June 1983

0 74820 08715 06

How you can tell



They're both IBM Personal Compute
And they're the same. Only different.

The system that's on the left is the ideal solution for a person who wants to be creative, efficient and improve his or her personal productivity.

So is the other one.

The system on the left incorporates the quality, the reliability and the technological excellence that have made IBM a computer leader for over 30 years.

So does the other one.

So which is which? Simple.

On the left is the IBM Personal Computer, starting with 64KB of user memory (expandable to 640KB) and two optional 5¼" diskette drives. It can easily satisfy your computing needs at the office, at home or in schoo

With 5 expansion slots it gives you room to grow. (You can even make it function like the computer shown on the right by adding an expansion unit that houses one or two 10-million-character fixed disk drives.)

This system can run most of the same software and accept most of the same IBM hardware as the computer on the right. And its price/performance is nothing less than remarkable.

The IBM Personal Computer

Established 1845 **SCIENTIFIC AMERICAN** June 1983 Volume 248 Number 6

ARTICLES

40 **PHYSICAL DISABILITY AND PUBLIC POLICY, by Gerben DeJong and Raymond Lifchez** Laws giving disabled people access to their environment remain largely unenforced.

50 **A VECTOR FOR INTRODUCING NEW GENES INTO PLANTS, by Mary-Dell Chilton**
In a natural form of genetic engineering plants are modified by a piece of DNA from a bacterium.

60 **GIANT VOLCANIC CALDERAS, by Peter Francis**
They are craters that remain after eruptions more violent than any in the span of human history.

86 **PEKING MAN, by Wu Rukang and Lin Shenglong**
The cave of Zhoukoudian has yielded a wealth of evidence on the life and times of *Homo erectus.*

96 **DARK MATTER IN SPIRAL GALAXIES, by Vera C. Rubin**
Much matter in galaxies emits no light, which bears on whether or not the universe is expanding.

110 **THE PHYSIOLOGICAL ECOLOGY OF WHALES AND PORPOISES, by John W. Kanwisher and Sam H. Ridgway** Their long dives are a part of their high-energy way of life.

122 **GEORG CANTOR AND THE ORIGINS OF TRANSFINITE SET THEORY, by Joseph W. Dauben** In the 1890's he succeeded in demonstrating that there is a hierarchy of infinities.

132 **THE SLIDE FASTENER, by Lewis Weiner**
Otherwise known as the zipper, it is the creation of many inventors going back to Elias Howe.

DEPARTMENTS

6 LETTERS
11 50 AND 100 YEARS AGO
12 THE AUTHORS
14 METAMAGICAL THEMAS
30 BOOKS
74 SCIENCE AND THE CITIZEN
146 THE AMATEUR SCIENTIST
154 BIBLIOGRAPHY

BOARD OF EDITORS Gerard Piel (Publisher), Dennis Flanagan (Editor), Brian P. Hayes (Associate Editor), Philip Morrison (Book Editor), John M. Benditt, Peter G. Brown, Michael Feirtag, Diana Lutz, Jonathan B. Piel, John Purcell, James T. Rogers, Armand Schwab, Jr., Joseph Wisnovsky

ART DEPARTMENT Samuel L. Howard (Art Director), Steven R. Black (Assistant Art Director), Ilil Arbel, Edward Bell

PRODUCTION DEPARTMENT Richard Sasso (Production Manager), Carol Hansen and Leo J. Petruzzi (Assistants to the Production Manager), Carol Eisler (Senior Production Associate), Karen O'Connor (Assistant Production Manager), Carol Albert, Lori Mogol, Martin O. K. Paul, Julio E. Xavier

COPY DEPARTMENT Sally Porter Jenks (Copy Chief), Debra Q. Bennett, Mary Knight, Dorothy R. Patterson

GENERAL MANAGER George S. Conn

ADVERTISING DIRECTOR C. John Kirby

CIRCULATION MANAGER William H. Yokel

SECRETARY Arlene Wright

SCIENTIFIC AMERICAN (ISSN 0036-8733), PUBLISHED MONTHLY BY SCIENTIFIC AMERICAN, INC., 415 MADISON AVENUE, NEW YORK, N.Y. 10017. COPYRIGHT © 1983 BY SCIENTIFIC AMERICAN, INC. ALL RIGHTS RESERVED. PRINTED IN THE U.S.A. NO PART OF THIS ISSUE MAY BE REPRODUCED BY ANY MECHANICAL, PHOTOGRAPHIC OR ELECTRONIC PROCESS, OR IN THE FORM OF A PHONOGRAPHIC RECORDING, NOR MAY IT BE STORED IN A RETRIEVAL SYSTEM, TRANSMITTED OR OTHERWISE COPIED FOR PUBLIC OR PRIVATE USE WITHOUT WRITTEN PERMISSION OF THE PUBLISHER. SECOND-CLASS POSTAGE PAID AT NEW YORK, N.Y., AND AT ADDITIONAL MAILING OFFICES. AUTHORIZED AS SECOND-CLASS MAIL BY THE POST OFFICE DEPARTMENT, OTTAWA, CANADA, AND FOR PAYMENT OF POSTAGE IN CASH. SUBSCRIPTION RATE: $24 PER YEAR IN THE U.S. AND ITS POSSESSIONS; $33 PER YEAR IN ALL OTHER COUNTRIES. POSTMASTER: SEND ADDRESS CHANGES TO SCIENTIFIC AMERICAN, 415 MADISON AVENUE, NEW YORK, N.Y. 10017.

SCIENTIFIC AMERICAN

CORRESPONDENCE

Offprints of more than 1,000 selected articles from earlier issues of this magazine, listed in an annual catalogue, are available at 75 cents each. Correspondence, orders and requests for the catalogue should be addressed to W. H. Freeman and Company, 4419 West 1980 South, Salt Lake City, UT 84121. Offprints adopted for classroom use may be ordered direct or through a college bookstore. Sets of 10 or more Offprints are collated by the publisher and are delivered as sets to bookstores.

Photocopying rights are hereby granted by Scientific American, Inc., to libraries and others registered with the Copyright Clearance Center (CCC) to photocopy articles in this issue of SCIENTIFIC AMERICAN for the flat fee of $2 per copy of each article or any part thereof. Such clearance does not extend to the photocopying of articles for promotion or other commercial purposes. Correspondence and payment should be addressed to Copyright Clearance Center, Inc., 21 Congress Street, Salem, MA 01970. Specify CCC Reference Number ISSN 0036–8733/83. $2.00 + 0.00.

Editorial correspondence should be addressed to The Editors, SCIENTIFIC AMERICAN, 415 Madison Avenue, New York, NY 10017. Manuscripts are submitted at the authors' risk and will not be returned unless they are accompanied by postage.

Advertising correspondence should be addressed to C. John Kirby, Advertising Director, SCIENTIFIC AMERICAN, 415 Madison Avenue, New York, NY 10017.

Subscription correspondence should be addressed to Subscription Manager, SCIENTIFIC AMERICAN, P.O. Box 5969, New York, NY 10017. The date of the last issue on your subscription is shown in the upper right-hand corner of each month's mailing label. For change of address notify us at least four weeks in advance. Please send your old address (if convenient, on a mailing label of a recent issue) as well as the new one.

Name

New Address

Street

City

State and ZIP

Old Address

Street

City

State and ZIP

THE COVER

The painting on the cover portrays a slide fastener, or zipper, as though it were being seen through a fairly powerful magnifying glass. The zipper is shown without its slider in order to depict the method of opening and closing, which take place inside the slider. This zipper has metal teeth, as all zippers did until improvements in plastics after World War II made it possible to manufacture zippers with nylon or polyester teeth. Each metal tooth has a projection on the top of one end and a corresponding pocket on the bottom. As the slider is raised its *V*-shaped internal neck holds the teeth at an angle that allows the projection of one tooth to slip under the pocket of the tooth on the opposite side, and then the *Y*-shaped flanges of the slider push the teeth together. The projection-pocket geometry is what holds the two sides of the zipper (the stringers) together when the zipper is closed. In opening the procedure is reversed; the slider's flanges hold the teeth at an angle that allows them to slip apart, and the neck separates them (see "The Slide Fastener," by Lewis Weiner, page 132). Each tooth is fastened to the fabric tape by legs clamped around the bead of the tape. The tape also serves for sewing the zipper to a garment.

THE ILLUSTRATIONS

Cover painting by Ted Lodigensky

Page	Source
14–19	Edward Bell
40	Ezra Stoller, © ESTO
42–49	Albert E. Miller
50	Robert and B. Gillian Turgeon, Cornell University (*top*); Mary-Dell Chilton (*bottom*)
52–57	Bunji Tagawa
58	Jacques Tempé
59	Bunji Tagawa
61	Natural Environment Research Council, Swindon, England, and Open University
62–63	Peter Francis
64–68	Tom Prentiss
70	Peter Francis
87	Carl Mydans
88–92	Patricia J. Wynne
93	Wu Rukang and Lin Shenglong, Institute of Vertebrate Paleontology and Paleoanthropology, Peking
94	Patricia J. Wynne
96–97	National Geographic Society–Palomar Observatory Sky Survey (*top left, middle center, bottom left and bottom center*), European Southern Observatory/ Science Research Council Southern Sky Atlas (*top center and top right*), Mount Wilson and Las Campanas Observatories, courtesy Allan R. Sandage (*middle right and bottom right*), Palomar Observatory, courtesy David Burstein (*middle left*)
98	Gabor Kiss
99–100	Vera C. Rubin, Carnegie Institution of Washington (*top*); Gabor Kiss (*bottom*)
101–105	Gabor Kiss
106	Vera C. Rubin, Carnegie Institution of Washington (*left*); Gabor Kiss (*right*)
108	Gabor Kiss
110	Sam H. Ridgway
112	Alan D. Iselin
114	U.S. Navy (*top*), Sam H. Ridgway (*bottom*)
115–119	Alan D. Iselin
123	Egbert Schneider
124–131	Jerome Kuhl
133–143	Ian Worpole
147	Frank R. Seufert
148–152	Michael Goodman

If you're not sure what size sweater to order, ask for Sharon.

She'll make sure you get the one that fits.

Sharon supervises the 90 friendly, well-informed operators who answer the phone when you dial the toll-free Lands' End number: 800-356-4444.

If she doesn't answer the phone herself, the person who does will be able to tell you what you want to know—about size, fit, material, color, availability—whatever. Including our unconditional, two-word guarantee:

GUARANTEED. PERIOD.

If you already have one of our catalogs, you will find it convenient to have it at hand when you call. If not, why not ask for one. Or, if you prefer, mail the coupon below. It's free.

Remember, you can call us toll-free any hour of the day or night. One of us will answer. We certainly don't want you to talk to a machine.

Lands' End Cable V-Neck

We've found considerable interest in the cotton sweaters we have offered for summer wear—on those evenings, for example, when dusk brings a chill with it. This particular model is our new Lands' End Cable V-Neck of 100% cotton. It's a sweater basic to any traditional wardrobe. Yet, it is a natural extension from basic crew, providing value. And a colorful one, too—in red, kelly, yellow, blue, lavender and natural. Discuss your selection with Sharon, why don't you?

LANDS' END
DIRECT MERCHANTS

of fine wool and cotton sweaters, Oxford button-down shirts, traditional dress clothing, snow wear, deck wear, original Lands' End soft luggage and a multitude of other quality goods from around the world.

☐ **Please send free catalog.**

Lands' End Dept. Q-05
Dodgeville, WI 53533

Name _____

Address _____

City _____

State _____ Zip _____

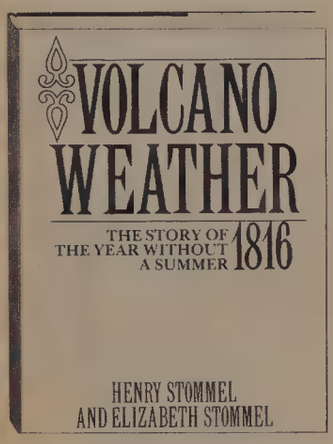
Or call Toll-free:

800-356-4444

(Except Alaska and Hawaii call 608-935-2788)

BIG BOOK from SMALL PRESS

You read it first in *Scientific American. Omni* is interviewing its authors. It's a spring alternate at Book-of-the-Month Club. Johnny Carson joked about it. *Booklist* said, "Knowledge vital for management of our ecology." *Country Journal* said, "A timely book."



No need to worry about "Big Brother" in 1984. Next year volcano weather will keep everybody quite occupied. The Stommels tell us what happened in 1816, a year after the volcano Tambora blew its top, and the extraordinary events that resulted in New England and Northern Europe.

VOLCANO WEATHER
is at bookstores now $15.00

SEVEN SEAS

Seven Seas Press Dept. 041
524 Thames Street
Newport, R.I. 02840

Play the finest

HARPSICHORDS & FORTEPIANOS

Build it yourself from one of our kits or let our experienced craftsmen build it for you.

Three Centuries of Harpsichord Making by Frank Hubbard (Harvard Univ. Press. 397 pgs. $20.00 postpaid)

For brochure send $1.00 to:
FRANK HUBBARD INCORPORATED
144-S Moody St., Waltham, MA 02154
(617) 894-3238

LETTERS

Sirs:

In the informative and entertaining article "The Footprints of Extinct Animals," by David J. Mossman and William A. S. Sarjeant [SCIENTIFIC AMERICAN, January], the authors cite as an example of paleoichnological misinterpretation of facts the 1834 discovery of Triassic-age tracks in the central German state of Thuringia that were originally attributed to an apelike or bearlike creature called *Chirotherium*. Later these tracks were shown to be those of a predominantly bipedal pseudosuchian reptile such as *Ticinosuchus*.

Mossman and Sarjeant also refer to sauropod tracks in the Cretaceous deposits at Paluxy Creek, Tex., discovered by the paleontologist Roland T. Bird in 1944. (Actually Bird published his first article on the Paluxy tracks in 1939.) They make no mention, however, of the alleged footprints of hominids alongside those of dinosaurs also found at Paluxy Creek in the 1930's. In view of current public debates between creationists and evolutionists it may be appropriate to set the record straight on this more famous misinterpretation.

Creationists would like to demonstrate that certain fossils are in the wrong order according to evolution theory so that they could prove the geological time scale is in error. One of their prime pieces of "evidence" in this regard has been the alleged existence of human footprints and the tracks of dinosaurs side by side in the Cretaceous limestone deposits of the Paluxy River near Glen Rose, Tex.

During the Great Depression of the 1930's some residents of Glen Rose had excavated dinosaur tracks and had fabricated other specimens, some "manlike," to sell to tourists. The fakery was disclosed in a 1939 article in *Natural History* by Bird. Subsequently other paleontologists who visited the Paluxy River site found tridactyl (three-toed) and sauropod dinosaur tracks and other Mesozoic plant and animal remains but no evidence of any Cenozoic fauna. Some of the dinosaur footprints were of such poor quality and/or so extensively eroded that they could be mistaken by lay people for "giant man prints." Some of these "man prints" have distinct claw marks protruding from what creationists call their "heel." Wann Langston, Jr., a Texas paleontologist, noted that these so-called human footprints have their "instep" along the outside edge rather than in the middle.

Some 10 years ago a film titled *Footprints in Stone* was made about the Paluxy River tracks by Films for Christ Association of Elmwood, Ill. Since the quality of the "human prints" was poor, the filmmakers decided to highlight the less obvious features such as "toes" and "sides" with shellac or oil. By stopping the film's motion Laurie R. Godfrey, a physical anthropologist at the University of Massachusetts at Amherst, has shown that the "man prints" all but disappear when the superposed outlines are eliminated. She observes that in other frames the "man print" was only a part of a larger impression or that the shellac seemed to connect erosional depressions. The film has been widely shown over the years with relatively little corresponding notice given to analyses of it from the scientific community.

According to Frederick Edwords, editor of the journal *Creation/Evolution:* "Last summer Dr. Godfrey led a research team to the site and discovered that the alleged human trackways showed no consistency. Many of the creationists' 'best' trackways featured irregular stride lengths and 'prints' that changed size and even direction with every step. The measurements she made will soon be published. Geologist Steven Schafersman discovered that worm burrows were largely responsible for the 'toes' on some of the prints."

Since *Scientific American* is widely read by both scientists (biological or otherwise) and nonscientists, it seems desirable to add this footnote to the article by Mossman and Sarjeant. Otherwise this particular myth may be perpetuated for yet another generation.

WILLIAM D. STANSFIELD

Professor, Biological Sciences
Department
California Polytechnic
State University
San Luis Obispo

EDITOR'S NOTE

In "The Law of the Sea," by Elisabeth Mann Borgese [SCIENTIFIC AMERICAN, March], it is stated: "Out of opposition to the Seabed Authority... the U.S., the U.S.S.R. and 15 other major industrial nations withheld their signatures from the Convention. These nations did, however, sign the Final Act of the Conference." The passage "the U.S., the U.S.S.R. and 15 other major industrial nations" should read "the U.S., the U.K. and certain other industrial nations." The error is not the responsibility of the author.

BREAKTHROUGH: FLOAT AN EYE TO STEADY ITS GAZE.

No matter how well protected mobile cameras or telescopes may be, the motion of the vehicle carrying them causes the lenses to jitter and shake. Images are blurred and imprecise.

We discovered a way to make a moving lens sit still. By using high-energy, rare earth magnets, we suspend an electro-optical system in a magnetic field—in effect, make it float. Untouched, it remains steady while its source of transportation vibrates. This "soft ride" assures crisp, unblurred vision for an observer or a sensor.

Army scouts call it the first major advance in mobile observation since binoculars were issued to the cavalry.

We're creating breakthroughs not only in optics but also in aerospace, health care, energy and information processing.

We're McDonnell Douglas.

MCDONNELL DOUGLAS

THE ALL-NEW 1984 FORD TEMPO

STYLE AND TECHNOLOGY IN TOTAL HARMONY

FORM FOLLOWS FUNCTION

In its creation of Tempo, Ford set out to design a car that would prove that a spacious interior, comfort and performance could be compatible with fuel efficiency. Ford engineers began with a sound design principle: Form follows function. First came complete new functional ideas. And then Tempo's pleasing form followed—naturally.

Available in 2-Door (above) and 4-Door (at right).

Designed by a world leader in aerodynamics.

Ford is an acknowledged world leader in aerodynamic design. And Ford's aerodynamic expertise is well expressed in the '84 Tempo.

Aerodynamic design can have a profound effect on the performance of any car.

No more than six horsepower is needed for Tempo to overcome air resistance at 50 mph. This results in less fuel consumed.

And impressive fuel economy figures are not the only result. The airflow is directed to reduce lift for improved stability and directional control.

The world's most advanced automotive computer.

If you've ever had your car's engine buck and stall on a cold morning...or choke up on a hot day...or had *any* irregular power output due to weather, temperature or altitude—you'll appreciate the capabilities of Tempo's new EEC-IV computer.

COMPUTER FUNCTIONS

- Pedal Position
- Fuel Air Mixture
- Spark Advance
- Intake-Air Temp.
- Engine Speed
- Crankshaft Position
- Engine Coolant Temp.
- Amount of O_2 in Exhaust Gas

EEC-IV (Electronic Engine Control) monitors and controls engine operation *precisely* and *instantly* under *any* conditions for optimum power output and fuel efficiency.

New Powertrain.

Developed specifically for Tempo, the 2300 HSC (High Swirl Combustion) four-cylinder engine works in concert with the EEC-IV computer.

This high-compression (9-to-1) engine generates 84 horsepower at 4400 rpm.† There's a surprisingly quick power response to even slight throttle pressure. Available with 4-speed, 5-speed or automatic transaxles.

High mileage.*

Although EPA mileage ratings were not available at the time of publication, Ford Engineering tests project 42 estimated highway and 28 estimated mpg.*

Ride and handling.

Tempo's front-wheel-drive design not only delivers added tire-to-road traction, but provides precise handling and tight cornering.

Tempo's new-design 4-wheel independent suspension permits

each wheel to absorb road shocks independently. This results in better handling and a smooth, quiet ride.

Computer refined interior.

Computer technology and designer ingenuity have joined forces to ensure the best use of Tempo's interior space. Its five-seat configuration allows for a comfortable combination of head, shoulder and legroom.

In fact, there is more rear head and legroom than in a Mercedes 300D.

Commitment to quality.

An example of Ford's commitment to quality is the closeness and uniformity of the body panel fits. You can compare Tempo's stringent standards to those set by any European Sports Sedan.

What's under Tempo's glistening paint? Three more full coats of tough acrylic enamel and two coats of primer. That's almost 10 pounds of protection. The final result: a tough, brilliant finish that is highly resistant to weather, chipping, cracking and peeling.

Pick up the Tempo of your life.

To obtain a brochure with more information on Ford's all-new sports sedan see your Ford Dealer or call 1-800-772-2100.

Get it together—Buckle up.

†Based on SAE standard J1349

*The above estimates are projected Ford ratings based on Ford Engineering's test data, and are expected to be very close to official EPA ratings. Use for comparison. Your mileage may differ depending on speed, distance and weather. Actual highway mileage and California ratings will probably be lower. See your Ford Dealer for a copy of the Gas Mileage Guide when available.

HAVE YOU DRIVEN A FORD...LATELY?

Ford

IF PERSONAL COMPUTERS ARE FOR EVERYBODY, HOW COME THEY'RE PRICED FOR NOBODY?

$1395* APPLE® IIe 64K

$999* TRS-80® III 16K

$1355* IBM® PC 64K

A personal computer is supposed to be a computer for persons. Not just wealthy persons. Or whiz-kid persons. Or privileged persons.

But person persons.

In other words, all the persons whom Apple, IBM, and Radio Shack seem to have forgotten about (including, most likely, you).

But that's okay. Because now you can get a high-powered home computer without taking out a second mortgage on your home.

It's the Commodore 64. We're not talking about a low-priced computer that can barely retain a phone number. We're talking about a memory of 64K. Which means it can perform tasks most other home computers can't. Including some of those that cost a lot more. (Take another look at the three computers above.)

By itself, the Commodore 64 is all the computer you'll ever need. Yet, if you do want to expand its capabilities some day, you can do so by adding a full complement of Commodore peripherals. Such as disk drives. Modems. And printers.

You can also play terrific games on the Commodore 64. Many of which will be far more challenging than those you could ever play on a game machine alone.

And as great as all this sounds, what's even greater-sounding is the price. It's hundreds of dollars less than that of our nearest competitor.

So while other companies are trying to take advantage of the computer revolution, it seems to us they're really taking advantage of something else:

Their customers.

*Manufacturers' suggested list prices as of March 20, 1983. Monitor included with TRS-80 III only. Commodore Business Machines, P.O. Box 500R, Conshohocken, PA 19428; Canada–3370 Pharmacy Avenue, Agincourt, Ont., Can. M1W 2K4.

commodore COMPUTER

Apple is a registered trademark of Apple Computer, Inc.
TRS-80 is a registered trademark of Tandy Corp. IBM is a registered trademark of International Business Machines Corp.

THE COMMODORE 64. UNDER $600.

You can't buy a better computer at twice the price.

50 AND 100 YEARS AGO

SCIENTIFIC AMERICAN

JUNE, 1933: "The disaster to the airship *Akron,* with its irreparable loss of brave and highly trained men, has shocked the entire country, and a vigorous but ill-advised clamor has arisen in the press and Congress for the discontinuance of all rigid-airship effort by the United States. It is to be hoped that not too hasty a decision will be made by Congress. The *Macon,* about to be completed, should certainly be put into service. Where long-range war operations are conducted by a fleet, the airship, with its enormous range and its ability to carry a number of fast scouting airplanes aloft, remains the most admirable source of information. Nothing can replace an airship for long-range scouting, and the Battle of Jutland would have had a far different termination had the British possessed a rigid airship of similar caliber to the one employed by the German navy."

"When Lt. Comdr. Frank M. Hawks wrote the specifications for the Texaco *Sky Chief,* he thought he was giving the manufacturer a knotty problem. He asked for an airplane that would carry fuel enough for at least 2,000 miles at well over 200 miles per hour, and he was insistent that the airplane have a landing speed not to exceed 70 m.p.h. Heretofore all fast airplanes had a landing speed in a ratio from 2½ or 3 to 1, and Commander Hawks was asking the manufacturer to improve that ratio to something like 4 to 1. That the problem had been solved by the Northrop Corporation was immediately obvious after the new *Sky Chief* had been taken up for a few flights. Tentative performance figures showed a top speed of 250 m.p.h. and a cruising speed of more than 200 m.p.h. Performance figures showed a landing speed of from 40 to 50 m.p.h., or a ratio of 5 to 1. The improvement of the ratio was obtained by the use of large wing flaps or air 'brakes.' The lower side of the trailing edge of the wing is split from the upper side for about 20 per cent of the wing chord. Attached to the wing proper by hinges, this flap is raised or lowered by a geared shaft actuated from the cockpit."

"The remarkable growth of contraceptive clinics in all the larger cities is a very clear indication that control of human reproduction by means of contraception has reached a degree of public approbation from which it will not recede in spite of all efforts to the contrary. One of the most frequently employed of these efforts is the announcement that promiscuity and an uncontrolled orgy of sex would follow the general dissemination of contraceptive information. Another measure is deliberately to foster confusion between contraception and abortion. Criminal abortion can more correctly be considered the direct result of the *absence* of contraception. Those who attempt to prevent the spread of contraceptive information are encouraging the continuation of a social system that makes abortion not only possible but also very common."

SCIENTIFIC AMERICAN

JUNE, 1883: "The great bridge that connects the two cities of New York and Brooklyn has been formally opened to public travel. The President of the United States, the Governor of New York and the mayors of the two cities took part in the ceremonies. On the New York side was a procession of the famous Seventh Regiment of 14 companies, preceded by a military band of 90 pieces, policemen mounted and on foot, acting as escort to the President, members of his Cabinet, Governor Cleveland and other notables, occupying 25 carriages. The President and guests walked across the bridge and were met by the Mayor of Brooklyn and other officials, and were escorted to the Sands Street bridge station by the Twenty-third Regiment. The salutes from the five naval vessels in the harbor and from Governors Island were an impressive feature in the 'pomp and circumstance' of the occasion. But the culmination of the display was reserved for the evening. It is doubtful whether there ever was, in this country at least, so magnificent an exhibition of pyrotechnics as that from the bridge. A levee held by the President and the Governor later in the evening at the Brooklyn Academy closed the public ceremonies, after which the bridge was thrown open to the public and thousands passed over from one city to the other, making a continuous procession during the entire night."

"The U.S. Board of Engineers for Fortifications has submitted a valuable report upon the practice in Europe with the heavy Armstrong, Woolwich and Krupp rifled guns. The conclusions of this report are that hammered steel in the late Spezia trials proved superior to any other material hitherto tested for armor plate. It seems probable that a hammered-steel plate, if equal in thickness to the belt armor of the *Inflexible* (22 inches), would stop the shot of the 100-ton gun (Armstrong) fired with its greatest practicable velocity. While the 12-inch rifled gun may prove a sufficient armament for the barbette batteries of our seacoast defenses against the lighter ironclads of foreign navies, iron turrets, armed with guns of 100 tons' weight at least, will be needed to meet the attack of armored ships of the latest construction. The 100-ton chambered Armstrong gun, throwing a projectile of a ton weight and fired with a charge of more than 700 pounds of powder, may be taken as a sample of the monstrous requirements of modern war."

"In the Brayton petroleum engine 10 gallons of unrefined petroleum are said to give a constant power equal to five horse power for 10 hours. Crude petroleum costs about six or possibly eight cents a gallon, making that the cost of five horse power per hour. The engine is run by the combustion of the vapor of petroleum united with atmospheric air under pressure. The motor, it is stated, has been fairly tried and appears to be constructed on reasonable principles."

"The growth of the open-hearth, or what is known as the Siemens-Martin, process of making steel has been no less remarkable than that of the Bessemer process. At present we have furnaces of a capacity from 15 to 25 tons, and by combining several furnaces single ingots weighing from 120 to 125 tons have been produced at Le Creusot in France. The world's production of open-hearth steel ingots for ship and boiler plates, propeller shafts, ordnance, wheels and axles, wire billets, armor plate, castings of various kinds and a multiplicity of other articles cannot have been less than 800,000 tons last year. The process has followed two somewhat dissimilar lines. In England iron ores of a pure quality are dissolved in a bath of pig iron, with the addition of only small quantities of scrap steel and iron. At Le Creusot large quantities of wrought iron are melted in the bath. The iron is of exceptional purity, containing less than .01 per cent of phosphorus and sulphur."

"There are not a great many points in commercial or trade matters as to which the average British manufacturer would be willing to admit the decided superiority of American concerns. But in one department those on the British side concede the unmistakable and pronounced superiority of the gentlemen on the other side. That department is advertising, which in the United States has been developed and systematized to an extent almost unknown in Britain. Advertising on a bold scale is looked upon as a *necessity* in the United States, whereas not a few British firms believe they confer a favor on the journals they patronize, if they advertise at all."

THE AUTHORS

GERBEN DEJONG and RAYMOND LIFCHEZ ("Physical Disability and Public Policy") bring complementary perspectives to the subject of their article. DeJong is senior research associate and assistant professor of rehabilitation medicine at the Tufts University School of Medicine. He attended Calvin College, receiving his B.A. in 1968. He went on to earn two master's degrees at the University of Michigan: one in public administration (1970) and one in economics and social policy (1972). His Ph.D. in public-policy studies (1980) is from Brandeis University. From 1972 to 1975 he was director of the social-services evaluation and analysis division of the Michigan Department of Social Services. In 1976 and 1977 he was project director of the Levinson Policy Institute at Brandeis. DeJong moved to Tufts in 1978. Lifchez is associate professor of architecture at the University of California at Berkeley and an architect in private practice. He was graduated from the University of Florida in 1955 with a bachelor's degree. He holds three master's degrees: in architecture (1957) and art history (1967) from Columbia University and in city planning from Berkeley (1972). From 1964 to 1969 he was a member of the faculty at Columbia, moving to Berkeley in 1971. He has worked for several architectural firms, including that of Eero Saarinen. He is the coauthor with Barbara S. Winslow of *Design for Independent Living: The Environment and Physically Disabled People* (University of California Press, 1982).

MARY-DELL CHILTON ("A Vector for Introducing New Genes into Plants") writes: "My undergraduate studies at the University of Illinois at Urbana-Champaign were in the chemistry curriculum. My graduate studies (also at Illinois) started in organic chemistry, but in taking courses for my two minors (biochemistry and microbiology) I found out about DNA and have had an unwavering interest in it ever since. As a postdoctoral student at the University of Washington I continued to work on the DNA of bacteria, mice and maize. I became a faculty member at the University of Washington in microbiology in 1970. In 1979 I accepted a position as associate professor of biology at Washington University in St. Louis, where interaction with scientists at Monsanto led my interests in the direction of the genetic engineering of higher plants. Last month my career took a new turn: I became executive director of agricultural biotechnology for CIBA-Geigy. I shall be responsible for the development of a new research unit in North Carolina, one of whose goals will be to exploit the gene-vector system described in this article."

PETER FRANCIS ("Giant Volcanic Calderas") was born in Zambia, which was then called Northern Rhodesia. He went to England to continue his education, attending the Imperial College of Science and Technology as an undergraduate and a graduate student. His Ph.D. in geology was granted by Imperial College in 1969. Since 1971 he has been a member of the faculty of the department of earth sciences at the Open University. Since 1981 he has served as senior visiting scientist at the Lunar and Planetary Institute in Houston, where he works on remote-sensing techniques for the investigation of volcanic terrain. Francis has led expeditions into volcanic territory in central America, South America and Africa.

WU RUKANG and LIN SHENGLONG ("Peking Man") are paleoanthropologists. Both are natives of China, and the anthropology of their country is their main scientific interest. Wu was graduated from the Central University of Chongqing with a B.S. in 1940. He moved to the U.S. in 1946, getting his M.S. (1947) and Ph.D. (1949) from Washington University in St. Louis. After earning his doctorate he returned to China, becoming director of the anatomy department at the Dalian Medical College. In 1956 he went to the Institute of Vertebrate Paleontology and Paleoanthropology of the Academy of Sciences of China as research fellow; he is currently vice-director of the institute. Wu is a member of the Academy of Sciences and president of the Association of Anatomical Sciences. Lin studied archaeology at Beijing University. He joined the faculty of the University of Inner Mongolia in 1959, leaving in 1962 to continue his graduate studies at the Institute of Vertebrate Paleontology and Paleoanthropology. He is currently assistant research fellow there.

VERA C. RUBIN ("Dark Matter in Spiral Galaxies") writes: "For the past 18 years I have been a staff member of the Department of Terrestrial Magnetism of the Carnegie Institution of Washington, where I work on the dynamics of galaxies. I am also adjunct staff member of the Mount Wilson and Las Campanas Observatories. I obtained degrees in astronomy from Vassar College (a B.A. in 1948), Cornell University (an M.A. in 1951) and Georgetown University (a Ph.D. in 1954). For my M.A. thesis I searched for a large-scale systemic motion of all the galaxies whose radial velocities were then known. For my Ph.D. thesis (under George Gamow) I studied the spatial distribution of galaxies. Both subjects, which are closely related to the work I currently do, were not widely studied again until the 1970's, when more data and large computers became available. My life is busy with travel to observatories, particularly Kitt Peak near Tucson and Cerro Tololo and Las Campanas in Chile."

JOHN W. KANWISHER and SAM H. RIDGWAY ("The Physiological Ecology of Whales and Porpoises") are physiologists who are interested in diving both by human beings and by marine mammals. Kanwisher was trained in biophysics. His Ph.D. in the subject was granted in 1951 by the University of Rochester. In 1952 at a program on Bimini Island he met the comparative physiologist P. F. Scholander. He writes that as a result of the meeting "my formal education in physics and engineering was instantly put aside for physiology." He has taught biological oceanography at Harvard University and the Massachusetts Institute of Technology. Kanwisher spends much of his time as a consulting engineer, and he has patented several devices to enable divers to remain longer below the surface. Ridgway is head of the Biomedical Division of the Naval Undersea Center. His B.S. (1958) and D.V.M. degrees are from Texas A&M University. He went on to get a doctorate in neurobiology from the University of Cambridge in 1973. He has served as research veterinarian at several institutions, including the U.S. Naval Missile Center, the University of Southern California and the University of California at Santa Barbara.

JOSEPH W. DAUBEN ("Georg Cantor and the Origins of Transfinite Set Theory") is professor of history and of the history of science at the Herbert H. Lehman College of the City University of New York. He got his bachelor's degree at Claremont College in 1966. His Ph.D., awarded in 1972, is from Harvard University. In 1973 and 1974 he was at the American Academy in Rome, where he studied mathematics, perspective and art in the Italian Renaissance. In 1977 and 1978 he was at the Institute for Advanced Study. In 1980 he was Mead-Swing Visiting Scholar at Oberlin College. In 1981 he was a visiting scholar at Harvard.

LEWIS WEINER ("The Slide Fastener") is a consulting engineer much of whose work is related to the subject of his article. Born in Czechoslovakia, he was educated in Prague at the Charles University. He came to the U.S. in 1941. He holds several patents for devices employed in the manufacture of slide fasteners.

We cut high student loan payments

(from an average of $151/month)

down to size.

(to an average of $98/month)

You can now reduce your student loan payments dramatically—by up to 50%—in the OPTIONS program at Sallie Mae.

Sallie Mae is a private corporation chartered by act of Congress in 1972 to make it easier for lenders to provide student loans. Now, we've been authorized to make student loans easier to pay off and we're doing just that with OPTIONS.

We simply pay off your existing NDSL, GSL, and FISL loans and create one new, 7%-interest, guaranteed student loan with no prepayment penalty. By choosing a longer payback period, payments that start low and rise gradually, or a combination of both, you end up with a single, lower monthly payment. The average reduction is from $151 to just $98 a month.

Loan Category	Typical Old Payment	Typical OPTIONS Payment
$5,001-$7,500	$ 96.00	$ 58.00
$7,501-$11,000	122.00	72.00
$11,001-$15,999	166.00	96.00
$16,000-or more	214.00	123.00

What OPTIONS can do for you. To get an idea of how much we can reduce your monthly payments today, estimate the amount you still owe on any NDSL, GSL, or FISL loans in good standing. Find the total on this chart and you'll see what we've done for borrowers like you.

With OPTIONS, you get to choose the payment plan that fits into *your* plans. When you get a student loan, most lenders tell you how much you'll pay, and for how long. With OPTIONS, *you* tell *us*. With OPTION 1, you can select a longer payback period, for a lower, unchanging monthly payment. With OPTION 2, payments start very low and increase gradually—like your income. OPTION 3 gives you graduated payments, too, but lets you pay off the loan faster.

If you have student loans to pay off, exercise your OPTIONS. OPTIONS is your opportunity to create the student loan payment plan that's right for you...to cut your student loan payments down to a size you can live with right now, when your income is probably lower than it will ever be again.

So if you owe more than $5,000 in NDSL, GSL, or FISL loans, don't miss this unique opportunity.

Let us give you more details, at no obligation. Thousands of student loan borrowers have already taken advantage of the OPTIONS program. To find out why, fill in and return the coupon. We'll send you a brochure that fully explains the OPTIONS open to you. Or, if you have any questions, call our toll-free 800 number.

Name_____

Address_____

City_____ State_____ Zip_____

SallieMae

OPTIONS

Mail to: B108
Student Loan Marketing Association,
OPTIONS, 1050 Thomas Jefferson St., N.W.
Washington, D.C. 20007. OR CALL:
800-821-7700 toll-free. Ask for OPTIONS.

©1983, Student Loan Marketing Association

METAMAGICAL THEMAS

The calculus of cooperation is tested through a lottery

by Douglas R. Hofstadter

And then one fine day out of the blue you get a letter from S. N. Platonia, the well-known Oklahoma oil trillionaire, mentioning that 20 leading rational thinkers have been selected to participate in a little game. "You are one of them," it says. "Each of you has a chance at winning a billion dollars, put up by the Platonia Institute for the Study of Human Irrationality. Here's how. If you wish, you may send a telegram with just your name on it to the Platonia Institute in Frogville, Okla. You may reverse the charges. If you reply within 48 hours, the billion is yours—unless there are two or more replies, in which case the prize is awarded to no one. And if no one replies, nothing will be awarded to anyone."

You have no way of knowing who the other 19 participants are; indeed, Platonia's letter states that the entire offer will be rescinded if it is detected that an attempt has been made by any participant to learn the identity of, or to establish contact with, any other participant. Moreover, it is a condition that the winner (if there is one) must agree in writing not to share the prize money with any other participant at any time in the future. This is to squelch any thoughts of cooperation, either before or after the prize is given out.

The brutal fact is that no one will know what anyone else is doing. Clearly everyone will want that billion. Clearly all will realize that if their name is *not* submitted, they will have no chance at all. Does this mean that 20 telegrams will arrive in Frogville, showing that even having transcendent levels of rationality—as you of course do—is of no help in such an excruciating situation?

This is the Platonia Dilemma, a little scenario I thought up recently in trying to get a better handle on the Prisoner's Dilemma, of which I wrote last month. The Prisoner's Dilemma can be formulated in terms resembling this dilemma. Imagine you receive a letter from the Platonia Institute telling you that you and just *one* other anonymous leading rational thinker have been selected for a modest cash giveaway. As before, both of you are asked to reply by telegram within 48 hours to the Platonia Institute, charges reversed. Your telegram is to include, apart from your name, just the word "cooperate" or the word "defect." If both of you wire "cooperate," you will each get $3. If both of you wire "defect," you will each get $1. If one of each word is received, the cooperator gets nothing and the defector gets $5.

What choice would you make? It would be nice if you both cooperated, so that you would each get $3, but does it not seem a bit unlikely? After all, who wants to get suckered by a nasty defector who gets $5 for being sneaky? Certainly not *you!* Therefore you would probably decide not to cooperate. It seems a regrettable but necessary choice. Of course, both of you, reasoning alike, come to the same conclusion. Hence you will both defect, and so you will get a mere dollar apiece. If, however, both of you had just been willing to take a chance, you could have got $3 apiece. What a pity!

It was my discomfort with this seemingly logical analysis of the "One-Round Prisoner's Dilemma" that led me to formulate the following letter, which I sent to 20 friends:

"Dear ——:

I am sending this letter by Special Delivery to 20 of you (namely various friends of mine around the country). I am proposing to all of you a one-round Prisoner's Dilemma game, the payoffs to be monetary (provided by *Scientific American*). It is very simple. Here is how it goes.

"Each of you is to give me a single letter: *C* or *D*, standing for 'cooperate' or 'defect.' This will be used as your move in a Prisoner's Dilemma with *each* of the 19 other players. The payoff matrix I am using for the Prisoner's Dilemma is the one enclosed numbered *1* [*see illustration on this page*].

"Thus if everyone sends in *C*, everyone will get $57, whereas if everyone sends in *D*, everyone will get $19. You can't lose! And of course anyone who sends in *D* will get at least as much as everyone else. If, for example, 11 people send in *C* and nine send in *D*, then the 11 *C*-ers will get $3 apiece from each of the other *C*-ers (making $30) and will get nothing from the *D*-ers. Therefore *C*-ers will get $30 each. The *D*-ers, in contrast, will pick up $5 apiece from each of the *C*-ers (making $55) and will get $1 from each of the other *D*-ers (making $8), for a grand total of $63. No matter what the distribution is, *D*-ers always do better than *C*-ers. Of course, the more *C*-ers there are, the better *everyone* will do!

"By the way, I should make it clear that in making your choice you should not aim to be the *winner* but simply to get as much *money* for yourself as possible. Thus you should be happier to get $30 (say as a result of saying *C* along with 10 others, even though the nine *D*-sayers get more than you) than to get $19 (by saying *D* along with everyone else, so that nobody 'beat' you). Furthermore, you are not supposed to think that at some later time you will meet with and be able to share the goods with your coparticipants. You are not aiming at maximizing the total number of dollars *Scientific American* shells out, only at maximizing the number of dollars that come to *you!*

"Of course, your hope is to be the *unique* defector, thereby really cleaning up: with 19 *C*-ers, you will get $95 and they will each get 18 times $3, namely $54. But why am I doing the multiplication or any of this figuring for you? You are very bright. So are the others. All about equally bright, I would say. Therefore all you need to do is tell me your choice. I want all the answers by telephone (call collect, please) *the day you receive this letter.*

"It is to be understood (it *almost* goes without saying, but not quite) that you are not to try to consult with others who

		Player *B*	
		Cooperates	Defects
Player *A*	Cooperates	(3,3)	(0,5)
	Defects	(5,0)	(1,1)

Where (*x*, *y*) means Player *A* gets *x* dollars and Player *B* gets *y* dollars.

A payoff matrix for the Prisoner's Dilemma

INTRODUCING THE PERFORMANCE SEDAN WITH A GENETIC ADVANTAGE.

Fifteen years ago, when big was beautiful and the auto industry was churning out ungainly V-8's, BMW introduced the 2002. A performance car.

Nine years later, big was out and the future was conceded to the tiny economy car. BMW introduced the 320i. A performance car.

Today the auto industry has embraced performance as the wave of the present. And BMW has introduced the 318i. A performance car—but one with an important difference.

It arrives with a heritage, as opposed to a history of recurring identity crises.

WHY A CAR'S PAST IS ITS FUTURE.

All cars, especially those aspiring to high performance, are products of engineering, technology—and genealogy.

The BMW 318i enjoys a substantial head start on all counts.

Its engine anticipated by years the current preference for fuel injection. The BMW 3-Series power plant serves as the basis for the BMW Formula One race engine—modified on the 318i by a new L-Jetronic system that generates even greater quickness and torque.

Its gearbox has been refined to match the greater torque, while retaining the characteristic precision that has inspired comparison to the linkage of fine timepieces.

Its new independent suspension is a virtual syllabus of lessons learned on racecourses, delivering handling and road holding Car and Driver judged "magnificent."

It offers a new graduated power-assist steering system, larger disc brakes, and such technological innovations as an electronic system that actually calculates when routine servicing is needed.

It is also among the automotive leaders in incongruity, having managed to boost fuel efficiency* at the same time it was improving its performance.

"THE FINEST QUALITY CAR AVAILABLE."

Inside, the 318i rebuts the myth that a noisy, bone-rattling ride under cramped conditions somehow certifies a vehicle as a true performance car.

It is roomier than its predecessors and well equipped with noise- and vibration-dampening measures. It also offers orthopedically-designed bucket seats and other functional amenities bearing the imprint of its racing lineage.

Its "overall ergonomics," summed up AutoWeek, "demonstrate why many consider the BMW the finest quality car available on the market today."

In short, the 318i continues and advances a 55-year commitment to performance. As today's fledgling performance cars will discover, there is no available shortcut to that experience.

If you see no need to discover the point along with them, we invite you to visit your nearest BMW dealer and test drive the 318i.

THE ULTIMATE DRIVING MACHINE.

*EPA-estimated 27 mpg, 38 highway. Figures are for comparison purposes only. Your actual mileage may vary, depending on speed, weather and trip length; actual highway mileage will most likely be lower. © 1983 BMW of North America, Inc. The BMW trademark and logo are registered. European Tourist Delivery can be arranged through your authorized U.S. BMW dealer.

ONLY

ATARI GIVES YOU SUCH EASILY ACCESSIBLE GRAPHICS AND SOUND CAPABILITIES IN A 64K COMPUTER.

ONLY *the ATARI 1200XL Home Computer combines 64K RAM computing power with outstanding graphics and sound capabilities that are so easy to access. You can create sophisticated graphics and sound effects using quick and simple commands. This allows you to write complex programs in much less time than with many other 64K computers.*

ONLY *the ATARI 1200XL Home Computer has a specially designed keyboard which is fully redefinable and features 8 programmable function keys controlling 16 functions. That's twice as many as the Commodore 64. In addition, the ATARI 1200XL keyboard locks and unlocks electronically to protect your programs.*

ONLY *the ATARI 1200XL Home Computer provides a revolutionary "help" key that performs two important functions. First, it can provide the user with helpful instructions on selected programs used with the computer. Second, the "help"*

THE NEW

A Warner Communications Company © 1983. Atari, Inc. All rights reserved. *®Trademark of VisiCorp.

key performs self-diagnosis—as requested by the user on the memory, audio visual circuitry and keyboard—to ensure that all components are in peak operating condition. And if you need additional help only Atari gives you a toll-free "help line" to call (800) 538-8543; in California 1-(800) 672-1404.

***ONLY** the ATARI 1200XL offers you a home computer compatible with virtually all ATARI Computer peripherals and software (compatibility that other new computers like the Commodore 64 don't offer). There are over 2,000 programs currently available for the ATARI 1200XL, including the new AtariWriter™ word processing system and VisiCalc.®**

***ONLY** Atari puts so much more in the new 1200XL Home Computer so you get so much more out of it.*

ATARI® 1200XL™

HOME COMPUTER

WINE & FOOD

Chardonnay

The [illegible] of California Chardonnay is upon us. Though its grape is a native of the Burgundy region of France, Chardonnay is now finding success in the red soil of the Sonoma Valley. Here it produces a wine of great world-class quality and sophistication.

WINE

After we have picked the grapes at optimum maturity and cold-fermented the wine to dryness, Chardonnay is then elevated to greatness by the judicious use of oak barrel aging.

Following an extensive two year barrel experiment, we selected French Nevers oak barrels to lend the desired character and complexity to our 1981 Proprietor's Reserve Chardonnay. The wine is [illegible] style that complements food. Ripe apple and citrus aromas are followed by rich varietal flavors especially balanced to the oak. The finish is classic Chardonnay, full and spicy.

FOOD

The full flavor of Chardonnay makes it an ideal complement to seafood, veal, turkey and chicken. Also, my wife Vicki and I enjoy our Chardonnay as an aperitif wine, served with sliced cheeses and fruit.

Because the marriage of wine and food is the most graceful expression of our art form, Vicki creates many recipes to enhance the unique character of our wines. One of my favorites is *Grilled Veal Steaks, Sonoma Style*. Seasoned with an oriental blend of pimento, parsley, watercress and garlic, the light and spicy flavors of this dish are a perfect match for Chardonnay.

If you would like to sample some of our Chardonnay recipes and learn more about wine, please write to the address below for your free brochure, "Wine and Food, Chardonnay." We would enjoy sharing them with you.

Discover Sebastiani Vineyards, a family winery in the Sonoma Valley.

Sam J. Sebastiani

Sam J. Sebastiani

Sebastiani Vineyards, Dept. C, P.O. Box AA, Sonoma, CA 95476

©1982 SEBASTIANI VINEYARDS

you guess have been asked to participate. In fact, please consult with no one at all. The purpose is to see what people will do on their own, in isolation. Finally, I would appreciate a short statement to go along with your choice, telling me *why* you made this particular one.

Yours,
Doug H.

"P.S. By the way, it may be helpful for you to imagine a related situation, the same as the present one except that you are told all the other players have *already* submitted their choice (say a week ago) and so you are the last. Now what do you do? Do you submit *D*, knowing full well that *their* answers are already committed to paper? Now suppose immediately after having submitted your *D* (or your *C*) in that circumstance, you are told that actually the others really *haven't* submitted their answers yet but that they are all doing it today. Would you retract your answer? Or what if you knew (or at least were told) that you were the *first* person being asked for an answer?

"And, one last thing to ponder, what would you do if the payoff matrix looked like the one enclosed labeled *2* [*see illustration on this page*]?"

I want to stress that this situation is not an *iterated* Prisoner's Dilemma (discussed in last month's column). It is a one-shot, multiperson Prisoner's Dilemma. There is no possibility of learning over a period of time anything about how the others are inclined to play. Therefore all the lessons described last month are not applicable here, since they depend on the situation's being iterated. All that each recipient of my letter could go on was this thought: There are 19 people out there, somewhat like me, all in the same boat, all grappling with the same issues as I am. In other words, there was nothing to rely on except pure reason.

I had much fun preparing the letter, deciding whom to send it to, anticipating the responses and then receiving them. Before I reveal the results I invite you to think about how you would play in such a contest. I should particularly like you to take seriously the assertion that everyone is very bright. In fact, let me expand on this idea, since I thought people perhaps did not really understand what I meant by it. Thus please consider the letter to contain the following clarifying paragraph:

"All of you are very rational people. Therefore I hardly need tell you that you are to make what you consider to be your maximally rational choice. In particular, feelings of morality, guilt, vague malaise and so on are to be disregarded. Reasoning alone (of course including reasoning about the others' reasoning) should be the basis of your decision. And please remember that everyone is being told this (including *this!*)!"

I was hoping for—and expecting—a particular outcome to the experiment. As I received the replies by telephone over the next few days I jotted down notes so that I had a record of what impelled various people to choose as they did. The result was not what I had expected; in fact, my friends "faked me out" considerably. We got into heated arguments about the "rational" thing to do, and everyone expressed much interest in the entire question.

I should like to quote some of the feelings expressed by my friends caught in this deliciously tricky situation. David Policansky opened his call tersely by saying, "O.K., Hofstadter, give me the $19!" Then he presented this argument for defecting: "What you're asking us to do, in effect, is to press one of two buttons, knowing nothing except that if we press button *D*, we'll get more than if we press button *C*. Therefore *D* is better. That is the essence of my argument. I defect."

Martin Gardner (yes, I asked him to participate) vividly expressed the mental turmoil he and many others went through. "Horrible dilemma," he said, "I really don't know what to do about it. If I wanted to maximize my money, I would choose *D* and expect that others would also; to maximize my satisfaction I'd choose *C* and hope other people would do the same (by the Kantian imperative). I don't know, though, how one should behave *rationally*. You get into endless regresses: 'If they all do *X*, then I should do *Y*, but then they'll anticipate that and do *Z*, and so....' You get trapped in an endless whirlpool. It's like Newcomb's paradox." So saying, Martin defected, with a sigh of regret.

In a way echoing Martin's feelings of confusion, Chris Morgan said: "More by intuition than by anything else, I'm coming to the conclusion that there's no way to deal with the paradoxes inherent in this situation. So I've decided to flip a coin, because I can't anticipate what the others are going to do. I think—but can't know—that they're all going to negate one another." Therefore while Chris was on the telephone he flipped a coin and "chose" to cooperate.

Sidney Nagel was very displeased with his conclusion. "I actually couldn't sleep last night because I was thinking about it. I *wanted* to be a cooperator, but I couldn't find any way of justifying it. The way I figured it, what I do isn't going to affect what anyone else does. I might as well consider that everything else is already fixed, in which case the best I can do for myself is to play a *D*."

Robert Axelrod, whose work proves the superiority of cooperative strategies in the *iterated* Prisoner's Dilemma,

		Player *B*	
		Cooperates	Defects
Player *A*	Cooperates	(3,3)	(0,50)
	Defects	(50,0)	(.01,.01)

A second payoff matrix for the Dilemma

saw no reason whatever to cooperate in a one-shot game and defected without compunction.

Dorothy Denning was brief: "I figure that if I defect, then I always do at least as well as I would have done if I had cooperated. And so I defect." She was one of the people who faked me out. Her husband, Peter, cooperated. I had predicted the reverse.

By now you have probably been counting. So far I have mentioned five defectors and two cooperators. Suppose you had been me and had got roughly a third of the calls, and they were five to two in favor of defection. Would you dare to extrapolate these statistics to roughly 14 to six? How in the world can seven individuals' choices have anything to do with 13 *other* individuals' choices? As Nagel said, certainly one choice cannot influence another (unless you believe in some kind of telepathic transmission, a possibility we shall discount here). What justification might there be for extrapolating these results?

Clearly any such justification would rely on the idea that people are in some sense "like" one another. It would rely on the idea that in complex and tricky decisions such as this one people will resort to a cluster of reasons, images, prejudices and vague notions, some of which will tend to push them one way, others of which the other way but whose overall impact will be to push a certain percentage of people toward one alternative and another percentage toward the other. In advance you cannot hope to predict what those percentages will be, but given a sample of people in the situation you can hope that their decisions will be "typical." Thus the notion that early returns running five to two in favor of defection can be extrapolated to a final result of 14 to six (or so) would be based on assuming that the seven people are acting "typically" for people confronted with these conflicting mental pressures.

The snag is that the mental pressures are not completely explicit; they are evoked by, but not totally spelled out by, the wording of the letter. Each person brings a unique set of images and associations to each word and concept, and it is the set of those images and associations that will collectively create in that

Bell System people are ready for new directions.

"WE'RE NOT JUST FACING CHANGE. WE'RE MAKING IT."

—Ian M. Ross, President, Bell Laboratories.

Ever since AT&T announced its planned divestiture of the 22 Bell telephone companies, I've been asked how the coming change will affect Bell Labs. Will we remain as creative and innovative? Will we remain intact? Will we retain those unique research and development capabilities responsible for our technological leadership and contributions to the nation?

My answer to each of those questions is yes.

I'm not suggesting that the challenges ahead are easy ones. But divestiture does give us more freedom to innovate, to apply our resources in new ways and in new markets. We'll be able to follow our technology wherever it leads. We intend to make the most of those opportunities.

Bell Labs as an institution will remain intact, although some of our employees have been reassigned to AT&T's American Bell subsidiary and others will be transferred to serve the divested companies. The major portion of our resources will continue to be devoted to developing new and improved systems and services for local telephone companies and their customers. We will also continue to provide the innovations

to assure that AT&T's interexchange network remains the most flexible and intelligent in the world.

Change is not new to us. We're an organization that creates change. We look forward to it. We thrive on it.

But we also face the future with the strengths of our past. We've earned an international reputation as a result of the depth and breadth of our work, our ability to attract and nurture outstanding scientific and engineering talent and our ability to inspire innovation. The wealth of creativity at Bell Labs never ceases to amaze me.

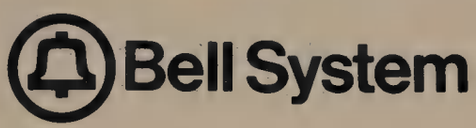
These same qualities are also responsible for the design and development of the world's finest telecommunications system. I'm proud of our role in delivering that system. I'm also proud that much of the technology of the Information Age is rooted in innovations conceived and developed by Bell Labs.

We're committed to making America's telephone system a total communications and information system. Our innovations will help AT&T maintain its leadership in these high technology fields. They will also provide the divested companies with the technologies needed to make them the true gateways to the Information Age.

We have a commitment to quality and reliability. In a world scrambling for quick markets and short-term profits, Bell Labs will continue to set standards of excellence. We believe that the increased competition that comes with divestiture will prove that quality and reliability will ultimately determine success.

As we have in the past, Bell Labs will continue to explore and expand the frontiers of research. To discover and develop new knowledge and new technology. To contribute to America's strength, economic growth and ability to meet any challenges from anyone.

At Bell Labs, we're not just facing change. We're making it.



person's mind a set of mental stresses like the set of stresses in the earth's crust in an earthquake zone. When people make their decision, you find out how all those stresses working in different directions add up, like a set of force vectors pushing in various directions with their strengths influenced by private or unmeasurable factors. The assumption that it is valid to extrapolate has to be based on the idea that everyone is alike inside, only with somewhat different weights attached to certain notions.

This way each person's decision can be likened to a geophysical experiment whose goal is to predict when and where an earthquake will occur. You set up a model of the earth's crust and you put in data representing your best understanding of the internal stresses. You know there are large uncertainties in your knowledge, and so you just have to choose what seem to be "reasonable" values for the variables. Therefore no single run of your simulation will have strong predictive power, but that's all right. You run it and you get a fault line telling you where the simulated earth's crust shifts. Then you go back and choose other values in the ranges of those variables and rerun the entire thing. If you rerun it repeatedly, eventually a pattern will emerge revealing where and how the crust is likely to shift and where it is rock solid.

This kind of simulation depends on an essential principle of statistics: the idea that when you let variables take on a few sample random values in their ranges, the overall outcome determined by a cluster of such variables will begin to emerge after a few trials and soon will give you an accurate model. You do not need to run your simulation millions of times to see valid trends emerging.

That is clearly the kind of assumption television networks make when they predict national election results on the basis of early returns from a few select towns in the East. Certainly they do not think that free will is any "freer" in the East than it is in the West, that whatever the East chooses to do the West will follow. It is just that the cluster of emotional and intellectual pressures on voters is much the same all over the nation. Obviously no one individual can be taken as representing the entire nation, but a well-selected group of residents of the East Coast can be assumed to be representative of the entire nation in terms of how much they are "pushed" by the various pressures of the election, so that their choices are likely to show general trends of the larger electorate.

Suppose it turned out that New Hampshire's Belknap County and California's Modoc County had produced, over many national elections, very similar results. Would it follow that one of the two counties had been exerting some kind of causal influence on the other? Would they have had to be in some kind of eerie cosmic resonance mediated by "sympathetic magic" for this to happen? Certainly not. All it takes is for the electorates of the two counties to be similar; then the pressures that determine how people vote will take over and automatically make the results come out similar. It is no more mysterious than the observation that a Belknap County schoolgirl and a Modoc County schoolboy will get the same answer when they are asked to divide 507 by 13: the laws of arithmetic are the same the world over, and they operate in the same way in remote minds without any need for "sympathetic magic."

All of this is elementary common sense; it should be the kind of thing any well-educated person understands automatically. And yet emotionally it cannot help but seem a little peculiar, since it flies in the face of the concept of free will and regards people's decisions as being simply the outcome of combinations of pressures with unknown values. On the other hand, perhaps it is better to look at decisions that way than it is to attribute them to free will, a philosophically murky notion at best.

That may have seemed like a digression about statistics and the question of individual actions v. the predictability of group behavior, but as a matter of fact it has plenty to do with the "correct action" to take in the dilemma of my letter. The question we were considering is: To what extent can what a *few* people do be taken as an indication of what *all* people will do? We can sharpen it: To what extent can what *one* person does be taken as an indication of what *all* people will do? The ultimate version of this question, stated in the first person, has a funny twist to it: To what extent does *my* choice tell me about the choices of the other participants?

You might think that each person is completely unique and therefore that no one can be relied on as a predictor of how other people will act, particularly in a situation involving an intense dilemma. There is more to the story, however. I tried to engineer the situation so that everyone would have the same image of it. In the dead center of the image was supposed to be the notion that everyone in the situation was using *reasoning alone*—including reasoning about the reasoning—to come to an answer.

Now, if reasoning dictates an answer, then everyone should independently come to that answer (just as the Belknap County schoolgirl and the Modoc County schoolboy would independently get 39 as their answer to the division problem). Seeing this fact is itself the critical step in the reasoning toward the correct answer, but unfortunately it eluded nearly everyone to whom I sent the letter. (That is why I came to wish I had included in the letter a paragraph stressing the rationality of the players.) Once you realize this fact it dawns on you that *either* all rational players will choose *D or* all rational players will choose *C*. That is the crux.

Any number of ideal rational thinkers faced with the same situation and suffering through similar throes of reasoning agony will necessarily come up with the same answer eventually, as long as reasoning alone is the ultimate justification for their conclusion. Otherwise reasoning would be subjective, not objective as arithmetic is. A conclusion reached by reasoning would be a matter of preference, not one of necessity. Now *some* people may believe this of reasoning, but rational thinkers understand that a valid argument must be *universally* compelling, otherwise it is simply not a valid argument.

If you grant this, then you are 90 percent of the way. All you need ask now is: "Since we are all going to submit the same letter, which one would be more logical? That is, which world is better for the *individual* rational thinker: one with all cooperators or one with all defectors?" The answer is immediate: "I get $57 if we all cooperate, $19 if we all defect. Clearly I prefer $57, hence cooperating is preferred by this particular rational thinker. Since I am typical, cooperating must be preferred by *all* rational thinkers. And so I'll cooperate." Another way of stating it, making it sound weirder, is this: "If I choose *C*, then everyone will choose *C*, so that I'll get $57. If I choose *D*, then everyone will choose *D*, so that I'll get $19. I'd rather have $57 than $19, and so I'll choose *C*. Then everyone will, and I'll get $57."

To many people this sounds like a belief in voodoo or sympathetic magic, a vision of a universe permeated by tenuous threads of synchronicity, conveying thoughts from mind to mind like pneumatic tubes carrying messages across Paris, and making people resonate to a secret harmony. Nothing could be further from the truth. This solution depends in no way on telepathy or bizarre forms of causality. It is just that the statement "I'll choose *C* and then everyone will," although entirely correct, is somewhat misleadingly phrased. It involves the word "choice," which is incompatible with the compelling quality of logic. Schoolchildren do not *choose* what 507 divided by 13 is; they figure it out. Analogously, my letter really did not allow choice; it demanded reasoning. Thus a better way to phrase the "voodoo" statement would be this: "If reasoning guides *me* to say *C*, then, since I am no different from anyone else as far as rational thinking is con-

THE TECHNOLOGY OF THE CENTURY.

THE 1983 BUICK CENTURY.

With today's technology, almost anything is possible. Computers that talk. Computers that teach. Electronics that put Wall Street right in your living room. And the Buick Century. A car so technologically advanced, it makes driving a pleasure instead of a chore.

The way Century accomplishes this is quite remarkable. There's an available electronic digital instrument panel, for example, that's a lot more than needles and dials. It tells you mph, km/h, fuel level, even how far you've traveled. All in computerized digital display.

Century's standard engine is electronically fuel injected. And connected to front-wheel drive. Its ride and handling are computer tuned. And all of this is surrounded by an aerodynamic shape that was refined in the wind tunnel, and further refined by computer.

Then there are, of course, the little things. Like an electronic tone for the ignition key and seat belt, instead of a buzzer. Why, Century is even built with robotic assembly techniques.

Happily, there remains one human quality that technology, even the technology of the Century, can never replace. The sheer pleasure of driving a Buick. And that's something you can experience at your Buick dealer now.

EST. HWY. 39 EPA EST. MPG 24

Use estimated MPG for comparison. Your mileage may differ depending on speed, distance, weather. Actual highway mileage lower. Some Buicks are equipped with engines produced by other GM divisions, subsidiaries, or affiliated companies worldwide. See your Buick dealer for details.

Official Car of the XXIIIrd Olympiad Los Angeles 1984

©1980 L.A. Olympic Committee ™

BUICK

Wouldn't you really rather have a Buick?

80 PROOF • CALVERT DISTILLERS CO., NYC

After the day's hassle and hustle, try a soothing change of pace.

John Jameson
Imported Irish Whiskey

cerned, it will guide everyone to say *C*."

The corresponding foray into the opposite world ("If I choose *D*, then everyone will choose *D*'") can be understood more clearly by likening it to a musing by the Belknap County schoolgirl before she divides: H'm, I'd guess that 13 into 507 is about 49—maybe 39. I see I'll have to calculate it out. But I know in advance that if I find it is 49, then sure as shooting that Modoc County kid will write 49 on his paper as well, and if I get 39 as my answer, then so will he. No secret transmissions are involved; all that is needed is the universality and uniformity of arithmetic. Similarly, the argument "Whatever I do everyone else will do" is simply a statement of faith that reasoning is universal, at least among rational thinkers, not an endorsement of any mystical kind of causality.

That analysis shows why you should cooperate even when the opaque envelopes containing the other players' answers are there on the table in front of you. Faced so concretely with this unalterable set of *C*'s and *D*'s, you might think: Whatever they have done, I am better off playing *D* than playing *C*, because certainly what I *now* choose can have no retroactive effect on what they chose. And so I defect. Such a thought, however, assumes that the logic now driving you to playing *D* has no connection or relation to the logic that earlier drove the other players to their decisions. But if you accept what was stated in the letter, then you must conclude that the decision you now make will be mirrored by the plays in the envelopes in front of you. If logic now coerces *you* to play *D*, it has *already* coerced the others to do the same, and for the same reasons; conversely, if logic coerces you to play *C*, it has also already coerced the others to do so.

Imagine there is a pile of envelopes on your desk, all containing other people's answers to the arithmetic problem "What is 507 divided by 13?" Having hurriedly calculated your answer, you are about to seal a sheet bearing "49" in your envelope when at the last minute you decide to check it. You discover your error and change the 4 to a 3. Do you at that moment visualize all the answers in the other envelopes suddenly pivoting on their heels and switching from 49 to 39? Of course not. You simply recognize that what is changed is your *image* of the contents of those envelopes, not the contents themselves. You first thought there were many 49's. You now think there are many 39's. It does not follow, however, that there was an instant when you thought: "They're all switching from 49 to 39." In fact, you would be crazy to think it.

It is similar with *D*'s and *C*'s. If at first you are inclined to play one way but on careful consideration you switch to the other way, the other players will not retroactively or synchronistically follow you. However, if you just give them credit for being able to see the logic you have seen, you must assume that their answers are the same as yours. In short, you are not going to be able to undercut them; you are simply in cahoots with them, like it or not! Either all *D*'s or all *C*'s. Take your pick.

Actually saying "Take your pick" is 100 percent misleading. It is *not* as if you could merely "pick" and then other people—even in the past—would magically follow suit! The point is that since you are going to be "choosing" by using what you believe is compelling *logic*, if you truly respect your logic's compelling quality, you would have to believe others would buy it as well, which means you are certainly *not* "just picking." In fact, the more convinced you are of what you are playing, the more certain you should be that others will also play (or have already played) the same way, and for the same reasons. This holds whether you play *C* or *D*, and it is the real core of the solution. Instead of being a paradox it is a self-reinforcing solution: a benign circle of logic.

If that still sounds like retrograde causality to you, consider this little tale, which may help to make it all make more sense. Suppose you and Jane are lovers of classical music. Over the years you have found that you have incredibly similar tastes in music—a remarkable coincidence. One day you find that two concerts are being given simultaneously in the city where you live. Both of them sound excellent to you, but Concert *A* simply cannot be missed, whereas Concert *B* is a strong temptation but one you will have to resist. Still, you are extremely curious about Concert *B*, because it features Zilenko Buznani, a violinist you have heard amazing things about.

At first you are disappointed, but then a thought flashes through your mind: Maybe I can at least get a firsthand report about Buznani's playing from Jane. Since she and I hear music through virtually the same ears, it would be almost as good as my going if *she* were to go. This is comforting for a moment, until it occurs to you that something is wrong. For the same reasons you would, Jane will insist on hearing Concert *A*. After all, she loves music in the same way you do—that's precisely why you want her to tell you about Concert *B*. The more you think Jane's taste is the same as yours, the more you want her to go to the other concert. But the more her taste is like yours, the less she will want to go to it!

The two of you are linked by a bond of common taste. And if it turns out that you differ enough in taste to disagree about which concert is better, then that

Over 300 manufacturers sell their microcomputers with Microsoft software.

There's a reason.

Powerful simplicity. That's the concept behind every software program we write. Powerful programs that let you spend more time thinking about the problem...and less time thinking about the computer. That concept has earned us the confidence of over 300 microcomputer manufacturers. Starting with the very first microcomputer you could buy. And today, Microsoft® languages, operating systems and applications software are running on well over a million microcomputers. Worldwide.

Made for each other. Manufacturers literally build their computers around Microsoft software. That's no exaggeration. In fact, Microsoft frequently participates in the initial design and development of microcomputers. That's a major reason why Microsoft software runs so well on the majority of the world's 8- and 16-bit systems.

User-oriented. Microsoft is the only software supplier to offer a full range of compatible operating systems, languages, utilities and applications programs. If you're not a computer expert, here's what that means to you: Better programs. Programs that are not only more powerful, but easier to learn and use.

Better tools. We can't buy your trust. We have to earn it. With better products. Tools you can easily use to solve complex problems. We started with the first BASIC for the first microcomputer. Today, we offer a broad range of proven tools for microcomputers. Including SoftCard™ and IBM® RAMCard,™ products that materially increase the capabilities of Apple® and IBM PC computers. Plus, business and management software such as Multiplan,™ the powerful electronic worksheet you can learn to use in just a few hours. Better tools. Because they're designed specifically for your computer.

Ask your Microsoft dealer. Most microcomputer manufacturers offer their systems with some Microsoft software, but you'll undoubtedly want more. Programs and languages that solve your specific problems. Programs for your industry, business or home. Programs that do more, yet ask less. Ask your computer software dealer for a demonstration. You'll see why more than 300 microcomputer manufacturers offer their systems with Microsoft software. The reason is powerful simplicity.

BETTER TOOLS FOR MICROCOMPUTERS

MICROSOFT™

MICROSOFT CORPORATION
10700 NORTHUP WAY
BELLEVUE, WASHINGTON 98004

Microsoft is a registered trademark, and Multiplan, SoftCard, RAMCard and the Microsoft logo are trademarks of Microsoft Corporation. Apple and the Apple logo are registered trademarks of Apple Computer, Inc. IBM and the IBM logo are registered trademarks of International Business Machines Corporation. The Radio Shack logo is a registered trademark of Radio Shack, a Division of Tandy Corporation. The Burroughs logo is a registered trademark of Burroughs Corporation. The Victor logo is a registered trademark of Victor Business Products. The Wang logo is a registered trademark of Wang Laboratories, Inc.

SOLID STATE.

The recent growth of "high-tech" industries has underscored the value of New Jersey's long term commitment to science and technology.

From Princeton's Institute for Advanced Studies and Rutgers, the State University, to the 28 state-approved vocational schools, the goal remains the same – to provide industry with the productive, quality workforce so critical in today's competitive marketplace.

Consider our educational resources, for example. New Jersey has five universities, 21 four-year colleges, 22 two-year colleges, a state college of engineering and architecture, plus a state medical and dental college.

Consider research and development. Nearly 10% of the nation's total R&D dollars are spent in New Jersey. 90% of these funds come from private industry. This makes New Jersey not only the leader in privately funded research, but the nation's leader in R&D overall.

Moreover, the State Government is dedicated to creating ideal conditions for business growth in New Jersey. In 1982, the N.J. Economic Development Authority channeled over $850 million in low interest financing to encourage new business development and expansion here.

New Jersey also offers many advantages in terms of location, people and lifestyles. The talent of our people is but one reason for saying: "New Jersey and You. Perfect Together."

To learn more about New Jersey contact:

Borden R. Putnam, Commissioner
Department of Commerce & Economic Development
CN 820, Trenton, NJ 08625
Phone (609) 292-2444

Thomas H. Kean, Governor

NJ & You
Perfect Together

will tend to make you lose interest in what Jane might report, since you no longer can trust her opinion as that of someone who hears music "through your ears." In other words, hoping she will choose Concert *B* is pointless, since it undermines your reasons for *caring* which concert she chooses!

The analogy is clear, I trust. Choosing *D* undermines your reasons for choosing it. To the extent that all of you in the game really *are* rational thinkers, you really will think along the same lines. And my letter was supposed to establish beyond doubt the notion that you are all "in sync," that is, to ensure you can depend on the others' thoughts to be rational, which is all you need.

Well, not quite. You need to depend not only on their being rational but also on their depending on everyone else's being rational, *and* on their depending on everyone to depend on everyone's being rational, and so on. A group of reasoners in this relationship to one another I call *superrational*. Superrational thinkers, by recursive definition, include in their calculations the fact that they are in a group of superrational thinkers. In this way they resemble elementary particles that are "renormalized."

A renormalized electron's way of interacting with, say, a renormalized photon takes into account that the photon's quantum-mechanical structure includes "virtual electrons" and that the electron's quantum-mechanical structure includes "virtual photons"; moreover, it takes into account that all these virtual particles (themselves renormalized) also interact with one another. An infinite cascade of possibilities ensues but is taken into account in one fell swoop by nature. Similarly, superrationality, or renormalized reasoning, involves seeing all the consequences of the fact that other renormalized reasoners are involved in the same situation—and doing so in a finite swoop rather than succumbing to an infinite regress of reasoning about reasoning about reasoning....

The answer I was hoping to receive from everyone was *C*. I was not so optimistic as to believe literally everyone would arrive at this conclusion, but I did expect a majority would, hence my dismay when the early returns strongly favored defection.

As more telephone calls came in I did get some *C*'s, but for the wrong reasons. Dan Dennett cooperated, saying "I would rather be the person who bought the Brooklyn Bridge than the person who sold it. Similarly, I'd feel better spending $3 gained by cooperating than $10 gained by defecting."

Charles Brenner, whom I had figured to be a surefire *D*, took me by surprise and *C*'d. When I asked him why, he candidly replied, "Because I don't want to go on record in a magazine of world-wide circulation as a defector." Very well. Know, world, that Charles Brenner is a cooperator!

Many people flirted with the idea that everyone would think "about the same" but did not take it seriously enough. Scott Buresh confided to me: "It was not an easy choice. I found myself in an oscillation mode, back and forth. I made an assumption, namely that everyone went through the same mental processes I did. I found myself wanting to cooperate roughly a third of the time. On the basis of that figure and the assumption that I was typical I figured about a third of the people would cooperate. Therefore I computed how much I stood to make in a field where six or seven people cooperate. It came out that if I were a *D*, I'd get about three times as much as I would if I were a *C*. So that I'd have to defect. Water seeks its own level, and I sank to the lower right-hand corner of the matrix." At this point I told Scott that so far a substantial majority had defected. He reacted swiftly: "Those rats—how can they all defect? It makes me so mad! I'm really disappointed in your friends, Doug."

So was I, when the final results were in: 14 people had defected and six had cooperated—exactly what the networks would have predicted. Defectors thus received $43 and cooperators got $15. I wonder what Dorothy is saying to Peter about now? I bet she is chuckling and remarking, "I told you I'd do better this way, didn't I?" Ah me, what can you do with people like that?

A striking aspect of Scott Buresh's answer is that in effect he treated his own brain as a simulation of other people's brains, and he ran the simulation long enough to get a sense of what a "typical person" would do. This is very much in the spirit of my letter. Having assessed what the statistics are likely to be, Scott then did a coolheaded calculation to maximize his profit, based on the assumption of six or seven cooperators. Of course, it came out in favor of defection. In fact, it would have done so no matter what the number of cooperators was. Any such calculation will always come out in favor of defection. As long as you feel your decision is *independent* of others' decisions, you should defect. What Scott failed to take into account was that coolheaded calculating people should take into account that coolheaded calculating people should take into account that coolheaded calculating people should take into account that....

That sounds awfully hard to take into account in a finite way, but actually it is the easiest thing in the world. All it means is that all these heavy-duty rational thinkers are going to see they are in a symmetrical situation, so that whatever reason dictates to one it will dictate to all. From this point on the process is simple. Which is better for an individual if it is a universal choice: cooperation or defection? That's all.

Actually it is not quite all, because I have swept one possibility under the rug, namely that maybe throwing a die could be better than making a deterministic choice. Like Chris Morgan, you might think the best thing to do is to choose *C* with probability p and *D* with probability 1 minus p. Chris arbitrarily let p be 1/2, but it could be any number between 0 and 1, where the two extremes represent defection and cooperation respectively. What value of p would be chosen by superrational players? It is easy to figure out in a two-person Prisoner's Dilemma, where you assume that both players use the same value of p. The expected earnings for each, as a function of p, come out to be $1 + 3p - p^2$, which grows monotonically as p increases from 0 to 1. Therefore the optimum value of p is 1, meaning certain cooperation. Where you have more players the computations get more complex, but the answer does not change: the expectation is always maximal when p is equal to 1. Hence this approach confirms the earlier one, which did not entertain probabilistic strategies.

Rolling a die to determine what you will do did not add anything new to the standard Prisoner's Dilemma, but what about the modified-matrix version I gave in the postscript to my letter? I shall let you figure that one out for yourself. And what about the Platonia Dilemma? There two things are clear: (1) if you decide not to send a telegram, your chances of winning are zero; (2) if everyone sends a telegram, your chances of winning are zero. If you believe what you choose will be the same as what everyone else chooses because you are all superrational, then neither of the alternatives is very appealing. With dice, however, a new option presents itself: roll a die with probability p of coming up "good" and send in your name if and only if "good" comes up.

Now imagine 20 people all doing this and figure out what value of p maximizes the likelihood of exactly *one* person getting the go-ahead. It turns out that it is $p = 1/20$ or, more generally, $p = 1/N$, where N is the number of participants. In the limit where N approaches infinity the chance that exactly one person will get the go-ahead is $1/e$, which is just under 37 percent. With 20 superrational players all throwing icosahedral dice the chance that *you* will come up the big winner is quite close to $1/(20e)$, which is a little less than 2 percent. That's not at all bad! Certainly it's a *lot* better than zero percent.

The objection many people raise is: "What if my roll comes up bad? Then why shouldn't I send in my name any-

LYNCHBURG
HARDWARE & GENERAL STORE®
47 Main St., Lynchburg, TN 37352

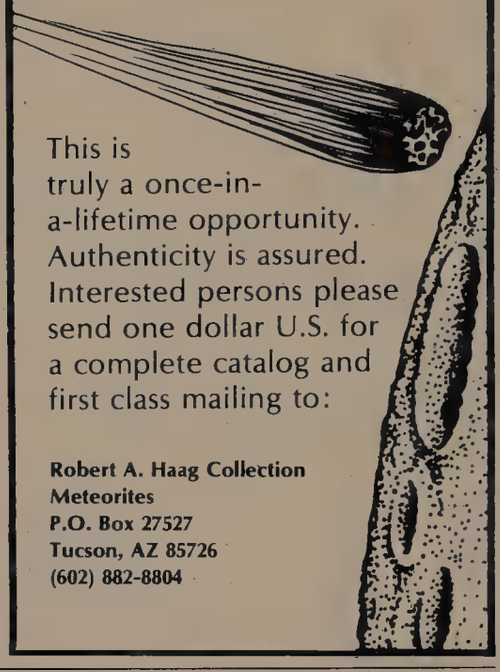
JACK DANIEL SUNGLASSES

These are some of the best looking sunglasses you'll ever find. "Jack Daniel's" is engraved right into the lens in gold coloring. We're offering these in traditional "aviator" styling. Please specify either smoke or bronze tint. My price of **$10.00** per pair includes postage and handling.

Send check, money order or [illegible] American Express, Diners Club, Visa or MasterCard, including all numbers and signature. (Add 6¾% sales tax for TN delivery.) For a free catalog, write to Eddie Swing at the above address. Telephone: 615-759-7184

Meteorites
Nothing on earth like it!

The Robert A. Haag meteorite collection is being offered for sale. Select iron and stone meteorites are available to astronomers, schools and private collections.



This is truly a once-in-a-lifetime opportunity. Authenticity is assured. Interested persons please send one dollar U.S. for a complete catalog and first class mailing to:

Robert A. Haag Collection
Meteorites
P.O. Box 27527
Tucson, AZ 85726
(602) 882-8804

AN UNDERSTANDING OF THE PAST FOR THE COMPREHENSION OF THE PRESENT

A RESEARCH CENTER SPECIALIZED IN PREHISTORIC ROCK ART

- Meetings, seminars and symposia
- *BCSP:* A World Journal of Rock Art Studies
- Publications on the origins and development of human culture
- Summer School: July -September
- International Archaeological Field Camps
- Exploration, expeditions and field work the world over
- World Inventory of Rock Art

CENTRO CAMUNO DI STUDI PREISTORICI

ROCK ART: 40.000 YEARS OF HUMAN INTELLECTUAL ADVENTURES

TAKE PART IN THE DISCOVERY OF YOUR INTELLECTUAL ROOTS

RETURN TO:
CENTRO CAMUNO DI STUDI PREISTORICI
25044 CAPO DI PONTE (Brescia) Italy
Phone: I - 0364/42091 - Telex I - 301504 Archeo

I wish to apply for membership in the Centro
Please send me further information and an application form

Name .....
Address .....
City ..... State ..... Zip .....

way? After all, if I fail to do so, I'll have no chance whatever of winning. I'll be no better off than if I had never rolled my die and had just voluntarily withdrawn." This objection seems convincing at first, but actually it is fallacious, being based on a misrepresentation of the meaning of "making a decision." A genuine decision to abide by the throw of a die means that you *must* abide by the throw of the die; if under certain circumstances you ignore the die and do something else, then you never made the decision you said you had made. Your decision is revealed by your actions, not by your words before acting.

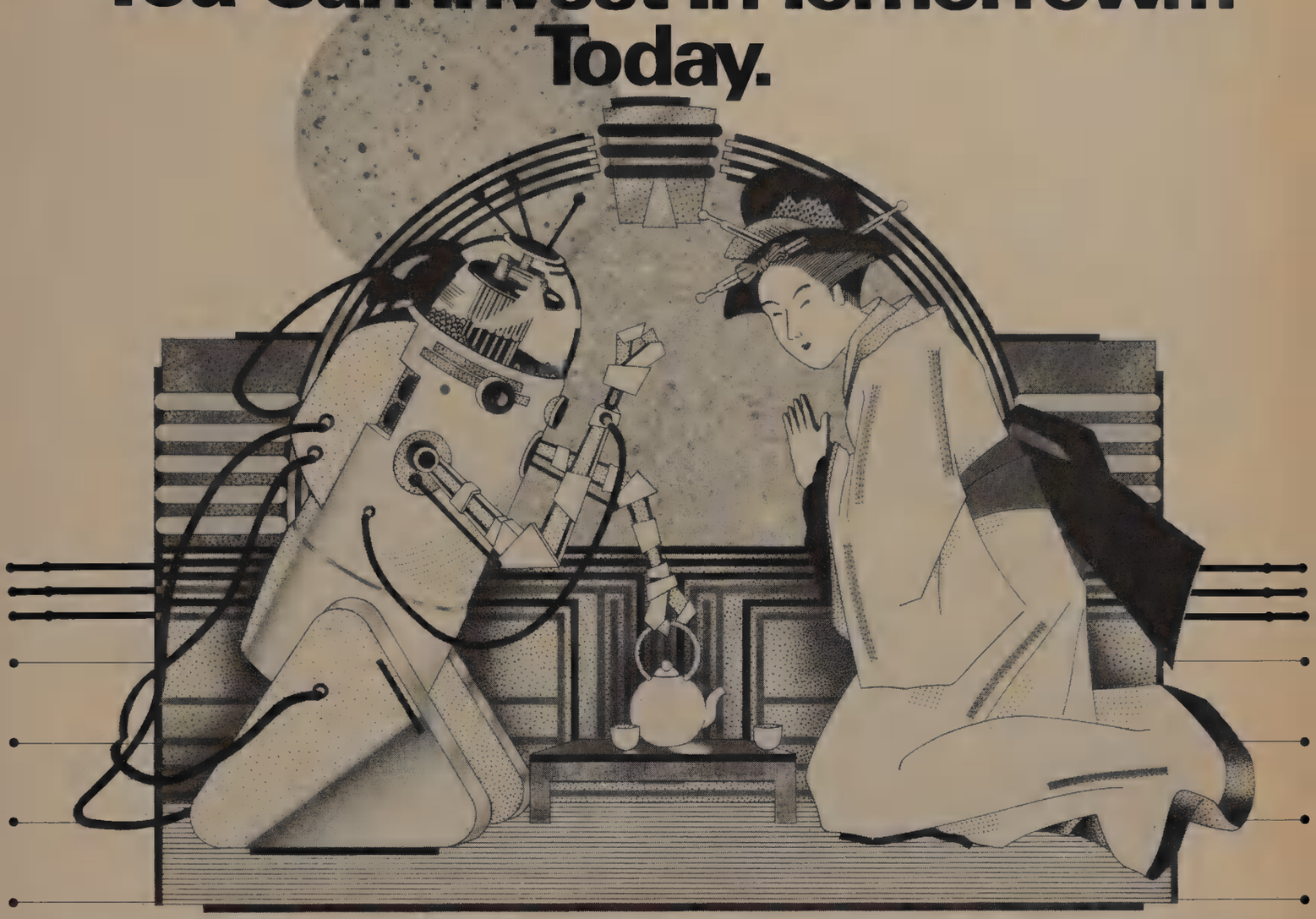
If you like the idea of rolling a die but fear your willpower may not be up to resisting the temptation to defect, imagine a third "Policansky button": this one says *R* (for "roll"), and if you press it, it rolls a die (perhaps simulated) and then irrevocably either sends your name or does not, depending on which way the die came up. This way you are never allowed to go back on your decision after the die is cast. Pushing *that* button is making a *genuine* decision to abide by the roll of a die. It would be easier on any ordinary human being to be thus shielded from the temptation, but any superrational player would have no trouble holding back after a bad roll.

This talk of holding back in the face of strong temptation brings me to the climax of this column: the announcement of a Luring Lottery open to all readers of *Scientific American*, and nonreaders too. The prize of the lottery is $1,000,000/*N*, where *N* is the number of entries submitted. Just think, if you are the only entrant (and if you submit only one entry), a cool million is yours! Perhaps, though, you doubt this will come about. It does seem a trifle iffy. If you would like to increase your chances of winning, you are encouraged to send in multiple entries without limit. Just send in one postcard per entry. If you send in 100 entries, you will have 100 times the chance of some poor soul who sends in only one entry. Come to think of it, why should you have to send in multiple entries separately? Just send *one* postcard with your name and address and a positive integer (telling how many entries you are making) to:

Luring Lottery
Scientific American
415 Madison Avenue
New York, N.Y. 10017

You will have the same chance of winning as if you had sent in that number of postcards. Illegible, incoherent, ill-specified or incomprehensible entries will be disqualified. Only entries received by 5:00 P.M. on June 30, 1983, will be considered. Good luck to you (but certainly not to any other reader of this column)!

Technology Is Creating Tomorrow. You Can Invest In Tomorrow... Today.



Modern technology has encircled the globe, and nowhere is it more apparent than in Japan. Now, ancient customs and traditions share eminence with amazing scientific accomplishments.

The Japanese excel in such fields as robotics, fiber optics, semiconductors, fermentation techniques, ceramics, factory and office automation. Great technological triumphs are also being recorded in other parts of the Far East, as well as in America and Europe.

These exciting developments call for new ways to invest, and World of Technology, Inc., a global high-technology Mutual Fund, makes it easy to participate in tomorrow's technological triumphs... today!

At least 50% of the Fund's portfolio is invested in U. S. companies, while the balance includes common stocks traded on major foreign exchanges. The Fund uses management talents of investment experts in London and Denver, selecting companies large and small... but all having strong potential for growth.

Today, send for more information concerning worldwide high-technology investment opportunities.

In continental U. S., Call Free
1·800·525·9274

In Colorado, Call Free
1·800·525·9769

A mutual fund investing in high technology worldwide.

World of Technology, Inc.

P. O. Box 1900
Denver, Colorado
80201

Name _____ Address _____

City _____ State _____ Zip _____ Phone _____

A prospectus containing more complete information, including all charges and expenses, will be sent upon request. Read it carefully before you invest or send money.

TSAO63G

BOOKS

The human primate, the land from the air, Hideki Yukawa's story, biological minerals

by Philip Morrison

THE HUMAN PRIMATE, by R. T. Passingham. W. H. Freeman and Company ($19.95). "If ever apes learn to read, I hope that they will find the secret" of how to become human in this book. So writes the author, an Oxford neurologist and psychologist whose research is on the monkeys, in an admirably judicious yet lighthearted review of the vexed question of just what it is that makes us extraordinary among our fellow primates. The book is scrupulously documented, its arguments set as far as is now possible on experimental results, not on mere plausible conjecture. Many issues are still moot.

One graph suffices to make the case that as between distinct species (certainly not by mere modest sexual dimorphism) biology is destiny. The graph shows the development of the brain in the human being and in the chimpanzee. The two curves grow steeply at about equal rates after conception; both reach a plateau. The growth of the chimpanzee brain, however, slows markedly just before birth; the growth of the human brain continues at the fetal rate for about two years after birth. The result is the sovereign difference: the human brain ends up about three times the size of the chimpanzee's.

The author takes pains to examine the crude measure of absolute brain weight, because it lies at the core of his own experimental researches. It is of course proportioned to the weight of the individual animal, but we share with the great apes brains that are large for our build compared with those of the monkeys, just as monkeys have brains large for their size compared with those of other mammals. The primates are a brainy bunch. So too are the sea mammals; we do not quite know why, although plainly their bodies are built on a quite different plan.

The human brain is set firmly in the primate pattern; a single curve fits the data well for the volume of neocortex (those little gray cells) as a function of the brain volume, one neat power law all the way from the lemurs to proud *Homo sapiens*. We have more of that powerful hardware, but only because the brain we have is bigger, and bigger primate brains regularly have relatively more neocortical volume. The number of cells in a cylinder of cortex of a given area is the same in all mammals; the big brains have a lower cell density to accommodate the more numerous interconnections but have thicker cortical layers. Even a couple of million years ago our australopithecine ancestors had a mean brain volume relatively larger than that of today's chimpanzees, although the brainiest individual chimpanzees exceed the mean of the fossil hominids in this measure. Only the rather subtle issue around the dominance of one cerebral hemisphere over language suggests that the chimpanzee brain may differ from ours in function more than it appears. In monkeys the removal of one cerebral hemisphere has little effect on the learning even of complex tasks.

The rest of the comparative anatomy of apes and human beings tells no very different story. The world appears much the same to chimpanzee and man. We share a poor sense of smell, own very sensitive pads on the fingers, see color by day, use stereoscopic clues for depth and are deaf to high audio frequencies. Like us, the chimpanzees can make sense of mirrors and photographs and can classify their world. Without words, chimpanzees and orangutans can pick out by feel at the first trial that one of two objects which matches a visible sample. The chimpanzee has good manual skills; all Old World monkeys and apes can in fact use the light precision grip, and chimpanzees can manipulate our tools well, to make a pot of tea or flick a switch. There is no doubt that we are better at it; no one expects the chimp to play Chopin presto. Walking is again in our favor; it is an old knack in the family, three or more million years old. A photograph shows a chimp upright on two feet with its arms full of bananas; chimpanzees know the usefulness of this posture, although a 50-yard run erect is as much as has been seen among them.

Such up-to-date comparative anatomy occupies about a third of the text; the rest is split between a similar discussion of comparative abilities and one of comparative social order. Matching and discrimination tests have been done with many mammals; monkeys and apes excel, even allowing for the fact that the testers have a rather evident primate bias. (The dolphin is doubtless a clever species, but there is no basis for saying more for it than that it may compare with some of the clever simians.) A small orangutan named Abang learned to strike a flake from a flint core and then use the flake to cut a string to release food from a box. No claim is made that Abang is like the first Paleolithic craftsman; the animal simply followed a careful demonstration.

Wild chimpanzees have been seen to use simple tools—the sticks, stones, twigs and leaves around them—to dig, pound, probe, sponge up, rake in and reach out, to throw and club and prop up what they would. Other apes and monkeys have also been seen to use many of these ready tools. The modification of tools—ape toolmaking, not simple use—is clear in captivity and is present, if rudimentary, in the wild. Chimpanzees have an entire repertory of tool using, but even a blue jay has been known to tear and crumple paper strips to help move food pellets closer to the side of its cage. Perhaps more to the point is social imitation, famous among the macaques of Japan and even commoner in chimpanzees. Again birds can show it. Social custom—particular grooming postures customary among one group of chimps and never seen in a group of matching habitat nearby—strikes a note of recognition.

For more than a century language has been seen as the Rubicon of human nature. A table here lists 13 identified vocalizations of chimpanzees in the wild: laughter, a pant-hoot for general approval, the call *wraa*, sounded when a human being or another predator is detected or when a dead chimpanzee is found. The vocabulary is touching, but it is no language. Passingham is inclined to credit the success of recent teaching of sign or symbol languages to the chimpanzees, although plainly the matter is not finally resolved. Children "have a thirst for language," but chimpanzees need careful tuition to pick up the simplest rules step by step, if indeed they can do it at all.

Of course, the seat of human culture is society, not the individual. The human family does not seem to hold a clear key; 90 percent of bird species are monogamous, and many primates too are tightly pair-bonded, in a rather random distribution among the groups of primates. The titi-monkey pair shown in a loving winding of furry tails are better as a symbol of marriage than any robin's nest; titis do seem to share plant food easily between adults and young. Chimpanzees share the meat they occasional-

DATSUN MAXIMA
OBVIOUS LUXURY...SURPRISING PERFORMANCE

Introducing the luxury sedan that was born to perform. The magnificent Datsun Maxima boasts a fuel-injected 2.4 liter six cylinder engine that merges surprising power with unsurpassed elegance.

Inside, virtually every conceivable creature comfort: cruise control on both the 5-speed and automatic, eight speaker stereo including cassette deck with Dolby, air conditioning, power windows, mirrors and door locks. A voice will tell you when your "Right door is open"—one of six vocal reminders. All this is standard.

The optional sunroof is power driven. The optional leather package includes a new digital control panel that monitors the engine's functions electronically.

The Datsun Maxima is available in four-door sedan or wagon, gas or diesel. Compare it feature by feature with any other car in its luxury class. Compare the performance. Then compare the price. Most importantly of all, remember it's built by the Nissan Motor Company, Ltd.—whose name stands for quality world-wide.

NISSAN
WE ARE DRIVEN
DATSUN

© 1983 NISSAN MOTOR CORPORATION IN U.S.A.

NISSAN

Up from the pits to

sensational sound

Because records get stuck, get stuck, get stuck...

Anyone who has heard a worn or scratched disc played on good audio equipment knows the problem of analog recordings. Record wear and damage cannot be avoided. Even brand-new discs are subject to wow, flutter, and distortion—the inherent drawbacks of conventional recording methods. To eliminate these difficulties, an entirely new approach to sound reproduction was developed: Compact Disc (CD) Digital Audio.

Hitachi introduces the future of sound

Drawing on a wealth of accumulated experience in laser, semiconductor, and audio technologies, Hitachi was able to come up with one of the world's first, completely digital sound systems for home use. Mirror-like CDs, just one fourth as large as standard LPs, are imprinted with coded series of reflective pits, each one just 0.9 × 0.5 microns in size, analogous to a speck of dust on the floor of an amphitheatre. In all, some 8 billion of these pits make up a single, one-hour disc side. During playback, they are scanned by an infrared laser beam, and the beam reading is reconverted into sound, with no wow, no flutter, and just 1/200 of the distortion heard on ordinary discs. And there's no wear at all because only laser light comes into contact with the specially coated CD surface.

Top performance from a compact, complex package

Within the Hitachi CD player itself are nine different large-scale integrated circuits (LSIs), each equivalent to some 32,000 transistors, yet taking up just a few square centimeters of space. Perhaps the most important of these is the 16-bit linear digital/analog converter which translates digital pickup signals into analog signals at a rate of 44,100 sound segments per second. It allows output virtually identical to the original sound, dynamic range and S/N ratios of over 90 dB, through standard amplifier/speaker systems.

Hitachi's DA-1000 digital audio disc player

Technical excellence is embodied in all Hitachi products

The development of a compact disc digital audio system is just one case demonstrating Hitachi's technological strength. You'll find other examples in all of Hitachi's products, from stereo amplifiers, tuners and tape decks to a whole host of high-quality electronic appliances. Our comprehensive technical expertise is your guarantee of convenience, easy operation and high reliability in every product that bears the Hitachi brand.

HITACHI

A World Leader in Technology

Inquiries to: Hitachi Sales Corporation of America, 401 West Artesia Blvd., Compton, Calif. 90220 Tel: (213) 537-8383

Northern Ontario/Canada

The Sounds of Silence

The dip and pull and ripple of the paddle, the whistle whirring of the reel, the echo of the loon. This is the place to refresh your senses far and away from the workaday world. Come North to 400,000 lakes, to wilderness excursions or fly-in fishing camps, or resorts of all sorts in our great out-of-doors. North to greet the dawn and the call of the wild. Phone TOLL FREE 1-800-828-8585 or from New York State 1-800-462-8404 or write: Ontario Travel, Queen's Park, Toronto M7A 2E5, Ontario/Canada.

ONTARIO
yours
to discover!®

ly take, but they form sex-divided bands rather than family groups; their transfer of meat is more of a "tolerated scrounging" than an example of customary sharing. There is evidence that some 1.5 million years ago hominid bands already used butchering sites; "whatever the basic social unit at that time, we must suppose that the two sexes were more dependent on each other in their economic life than is true of chimpanzees today."

Cooperation and indeed competition play their roles in the social groupings of animals. Monkey groups sometimes fight battles, but only human beings sustain war. A careful account makes it plain that human warfare is not inbuilt aggression, as the ethologist Konrad Lorenz seemed to argue, or an analogue of territoriality or of dominance hierarchies. It is not the "beast within" we need to blame. Man is the only warring primate because he alone is able to make war; he is too clever for his own good. The author considers both the genetic and the cultural aspects of the human heritage at some length; he concludes that the current controversies are premature in the absence of adequate evidence. Both nature and nurture affect behavior. These days more people might be persuaded by a trendier phrase: the need for both hardware and software. The profound software changes of human prehistory in the absence of demonstrable genetic change at any such breakneck rate, however, place a heavy burden on sociobiological explanations.

The future evolutionary task of the chimpanzees is clear. "They must expand their brains according to the primate pattern" until they can fluently form and transmit propositions about the physical and the social world, even in the sensory absence of the topical stimulus. A novel social order will follow, a growing culture, and the apes will be on their way. What we can say today about that path may be elementary, but it does not seem to present an utter mystery. Over some 10 million years of change a threshold was crossed, by brain expansion rather than by reconstruction, and a rich and ready code arose. "Of course, it is possible that, when faced with the option of becoming a man, a rational chimpanzee would take one look at human society and turn down the offer."

BRITAIN FROM THE AIR, by Bernard Stonehouse. Crown Publishers, Inc. ($30). THE EXTRAORDINARY LANDSCAPE: AERIAL PHOTOGRAPHS OF AMERICA, by William Garnett. New York Graphic Society, Little, Brown and Company ($60). The cameramen of Aerofilms, Ltd., began in about 1919 (there is a takeoff picture of a stick-and-wire biplane to prove it) to assemble their unmatched collection of commercial air photographs of the island of Britain (and a little outside it). Stonehouse, an experienced pilot and a well-known ecologist, has lovingly compiled from that overflowing archive a selection of its photographs in color, mainly from flights over the past 10 years. He has prepared a brief description of each photograph, provided a map of the implied wanderings from the island of Jersey to the far Scottish Highlands and displayed the photographs on the large pages of this album. About seven of every eight photographs focus, speaking very roughly, on the work of British hands 1,000 feet below; the rest show a landscape that is mainly nature's.

Here, for instance, are the famous bridges of Britain, monuments in the history of modern technology. They are shown of course as they lie in the countryside, an essential view not so familiar to students of engineering from the books. First see Coalbrookdale, that delicate masterpiece of Abraham Darby's cast-iron foundry, spanning the narrow blue Severn for 200 years now. The first modern suspension bridge, designed by Thomas Telford, soars over the Menai Strait in northern Wales, modern steel cables bearing the load carried until quite recent times by the original huge links of wrought-iron chain. Close by is little Conwy, where the same granite headland that bears a fine 13th-century castle also receives three early Victorian wonders: Telford's suspension bridge for the coach road and the twin rectangular iron tubes designed by Robert Stephenson that still enclose and support the railway, one track within each.

Here too is a span of Isambard Kingdom Brunel's, a chain-link flung high above the sluggish Avon below Bristol, "completed with second-hand chains." The cantilevers of the Firth of Forth hold out their powerful hands to bear the trains, and behind them the elegant new cable-suspension highway bridge is mistily visible. The Severn cable bridge spans wide, windy waters near the sea, and the half-dozen busy crossings of central London, low arches and high towers, appear repeatedly.

An older tale is told in other structures. The export industry of Norfolk about 4,000 years back was flint tools. At Grime's Graves the field is pocked with long-overgrown spoil heaps and long-filled pits, where once Neolithic miners toiled with reindeer-antler picks to win the nodules of flint 20 feet down in the chalk. Here are the Avebury rings, near Stonehenge on Salisbury Plain, best seen as a concentric whole from the air, except that the viewer can then only envy the people whose parked cars testify to their pleasant hour at the country pub near the center. Scarred to the white chalk just below the roots of rich green pasture on Berkshire Downs is the giant silhouette of a "wild horse-goddess," with an Iron Age fort close by. Roman occupation is marked by a stone garrison fort at Housesteads along Hadrian's Wall, where a battalion of legionaries once served.

Here are the centers of learning, the University of Oxford lying grandly around the rotunda of the Radcliffe Camera, Cambridge seen from across the Cam, with the chapel of King's College at the center, and Eton College, framed a little too tightly to display more than a corner of the aphoristic playing fields. There are steel mills and mining towns, and one enormous nonnuclear power station, magically transformed almost into a Turner by billows of mist and the vapor columns rising from the eight drums of its cooling towers. British foibles are not avoided: endearing architectural follies, cricket and the crowd leaving a Cup Final at Wembley Stadium. There is a multiple pileup on the M1 road, a mortal disaster among dozens of cars and lorries in the fog somehow more devastating in Bedfordshire than in Los Angeles.

Among the images of nature the cliffs of Albion are strikingly shown, the green downs edged by 500 feet of white chalk, the sea below completely lost in mist. This is a striking set of documents, craftsmanlike rather than artful, but one image crosses that line. It shows the Pool of London at sunset, ships, barges and bridges silhouetted strongly against a winding golden Thames, the city an almost unmarked black. It would evocatively illustrate an edition of Conrad.

William Garnett, the author of the other book of aerial photographs, is no company but an engaged artist, one man with a million miles of solo flight in a light plane behind him, camera in one hand, controls in the other. His quarry is the beauty of the land. A few people appear in boats, as the lone fisherman adds melancholy scale to the landscape of the Chinese painters. The only works of humankind seen often here are the boldly contoured furrows of the tractor-driving plowman, along with field boundaries and windrows.

Here also there are two classes of images, both beautiful. They are superbly reproduced in this book, a triumph of American bookmakers from coast to coast. One class is a landscape where scale is easy to read: the inverted cone of a volcanic gas vent newly melted into glacier ice and snowpack high on Mount Baker or 100 miles of Mississippi oxbows glowing in the low sun. (Garnett flies and shoots at altitudes from 100 feet up to three or four miles.) In the other class the scale is covert, enigmatic. The picture is largely texture, a play of

form and color that may remind you of the land but might be the abstract canvas of a painter. The tufts of sunstruck brown kelp float in the blue sea. The eroded butte is a sculptured surface, the blue sky mirrored in the branching ravines filled with snow, the steeper slopes lacing the ravines with a chiseled gray-gold smoothness. Sometimes it is the clouds that bear the tale, sometimes it is the surf breaking on an unseen slope. The flooded fields of California rice glow deep sky blue, and across them the low dikes trace a contour map in sun gold.

Ansel Adams writes of these photographs as revelation; they disclose patterns we ordinarily miss. Garnett recalls 36 years of piloting, often intensely hard, dangerous work. He praises his tools: "Deep appreciation goes to all those unknown people who carefully engineered my plane,... crafted the parts,... gave my craft its rated strength and the integrity to serve... through unrelenting, pounding turbulence, and yet to respond with... ballerinalike movement in smooth air. Yes, to all you people at Cessna Aircraft, McCauley Propeller and Continental Engine—I love you for your care. Thank you for three decades of safe, beautiful, privileged flight." The opening photograph, by the artist's son, shows the silvery 1955 Cessna 170-B in flight; the closing image displays "the shadow of my plane on a cloud, completely encircled by a rainbow." It is a token for this happy union of technology and sensibility, in a life of devotion to an antique and spacious vision that became real only in our times.

HIDEKI YUKAWA: "TABIBITO" (THE TRAVELER), translated by L. M. Brown and R. Yoshida. World Scientific Publishing Co., Singapore. Distributed in the U.S. by Heyden & Son, Inc., 247 South 41 Street, Philadelphia, Pa. 19104 ($33). "Hideki Ogawa was born in 1907 at old Tokyo's Ichibei-cho Azabu. The house smelled of plum blossoms each spring.... I cannot remember the house of my birth, not even the plum blossoms there. I know them only through my mother, but in my mind I know they are extremely beautiful. (Perhaps my subconscious tries to beautify my birth.)" So began the physicist Yukawa when at 50, a decade after his name was the first from the Far East to enter the long list of Nobel prizewinners, he wrote this candid, gentle, reflective autobiography of his youth. (Like his father before him he had adopted his wife's family name on marriage, a custom in Japan when there is no son to continue the name.) The memoir ends in November, 1934, with his famous paper in English, the birth of the physics of the hidden transient particles of nature. (Enrico Fermi's neutrino theory of the same year probably shares that proud claim.) Yukawa was then a young lecturer in theoretical physics at Osaka Imperial University. He finished the paper quickly because "my wife kept urging me, 'Please write the English paper and show it to the world.'"

The flood of post-Meiji modernization in Japan runs brimful through the book. Hideki's paternal grandfather taught Chinese learning to the children of the feudal clan of Tanabē. Hideki's father, inclined to a bookish life, was drawn by a chance at the New Knowledge, and made his way to the university in Tokyo. The terrible earthquake of 1891 near Nagoya, not far from his home countryside, gripped his imagination. He visited the scene and came away in awe. "It was then that his life began to center around geology." He became a well-known field geologist. After a long government-sponsored trip to China to examine its minerals on the eve of the Japanese expansion, he went to Kyoto University, eventually to become dean. The infant Hideki grew up in Kyoto with two sisters and four brothers. Three of the sons became university professors; the other was killed in World War II.

Hideki was a shy, bookish, silent, somewhat lonely child. He was distant from his parents. His father did not intend to send this enigmatic boy to university; a technical school seemed good enough. But his mother's pleas and the intervention of the insightful principal of his school ("A boy with Hideki's potential is very rare") opened the path. Hideki knew he was mathematically gifted. An older brother brought home from the middle school the wonderful formula for summing a long arithmetic progression all at once. Younger brother had already worked out the method on his own, innocent of algebra and its notation. Mother realized it and "began to smile in delight."

The narrative is familiar; worldwide human unity outweighs the powerful old culture. This is a Japan as vibrant as it is constrained. Einstein visits, and all Kyoto—except Hideki—turns out to see and hear him. Dostoevski and de Mille are in everyone's mind; Hideki, still in his wide pleated trousers, has a nickname borrowed from Tolstoi, perhaps meaning Ivan the Fool. He is active in baseball, but he walks apart from even his sisters. He is taught at home the Chinese characters from the classics by rote without understanding, he swims in the traditional way (endurance for many hours with no account of distance) and the high school play is storytelling of the third shogun, a performance full of swordsmen and dependent on elegant female impersonators. (Hideki refused such a part.) His family was a strict one; playfulness was rare.

He was aimed at mathematics, pretty surely, until he hit a martinet of a high school teacher. The man marked as wrong a correct proof that differed from the one given in class; Hideki could not "dedicate his life" to a subject like that. A bad draftsman (although an able calligrapher), he could not hope for engineering. His geologist father made an effort to lead the boy toward geology, but Hideki's mind was a logical one, with little power for observation or memorization. He could not finish the big textbook in English his father produced out of the library at home. Physics it was to be; two German books on the old quantum theory (Reiche and Planck) made up his young mind.

Here is his undergraduate physics class photograph at Kyoto; visiting professor Otto Laporte of Michigan is there, and so is Hideki's close friend Sinitiro Tomonaga, who was to win a Nobel prize a few years after him. A young graduate student, Hideki is attracted by the "spirit of Copenhagen" in the lectures of a paternal Yoshio Nishina on quantum mechanics. That attitude is hard to describe, but "it is certain that it had much in common with the spirit of generosity."

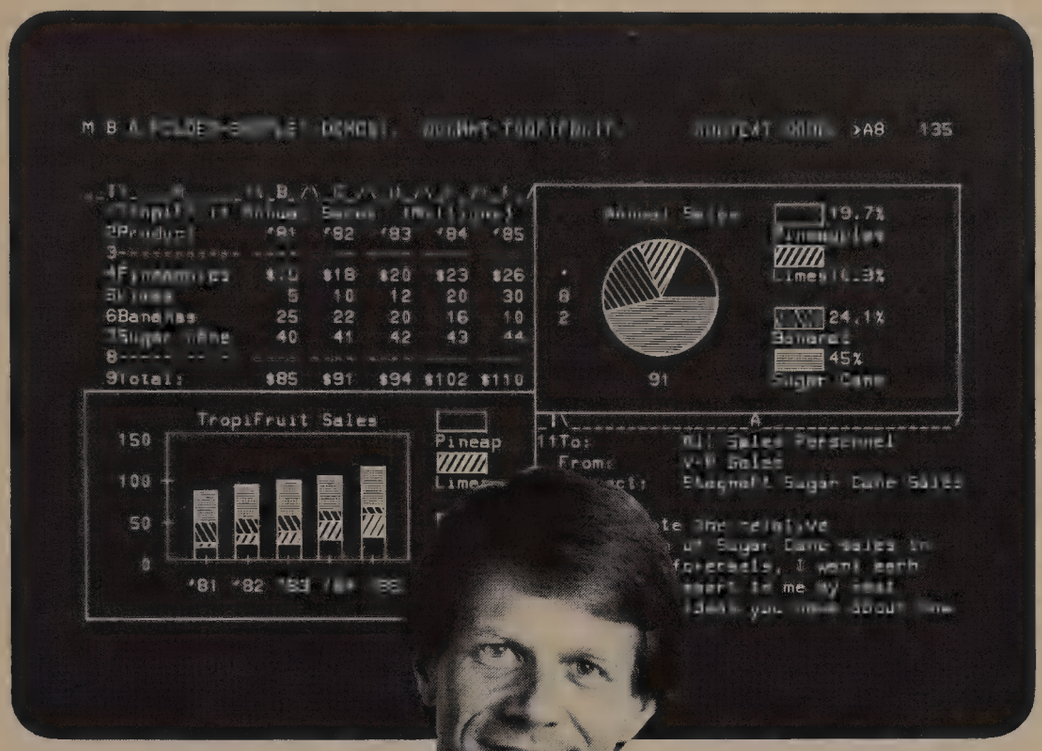
An arranged marriage is charmingly described; it was a lifelong success. The new Ph.D. finds his way, at first only part-time and then in a real teaching job, to the new, lively University of Osaka. There is a physicist-president there, and Yukawa is recruited by Hidetsugu Yagi, inventor of the antenna named after him. ("I developed a deep trust in this man.") No gifted teacher, Yukawa spoke to the blackboard, and his soft unemphatic speech invited sleep.

All the time he was seized with the refractory question of the nuclear forces. His first original research paper treated the magnetic interaction between the atomic electron and the nucleus, but Fermi beats him to the spectroscopic answer and goes further. Once again Yukawa tries for the nuclear forces; this time it is an effort to use the electron and the neutrino, where once more Fermi far off in Rome has blazed a path only dimly seen in Osaka. Finally, in a new house in the hills, a 10-mile commute to the dusty and busy streets of industrial Osaka, he is forced to spend a few days at home. Sumi, his wife, is confined with the birth of their second child, and stormy fall weather makes it hard to travel among the fallen trees.

One night, unable to sleep, the new infant a few days old, the crucial point comes to him. The short range of nuclear forces demands a carrier of the force field with a mass of 200 electrons. No one could easily observe such a particle. It is present but hidden, if his calculations are right. That day in the meeting of their lively nuclear research group he

IBM
2½ minutes

HP
30 seconds



Hewlett-Packard's new personal computer is up to 5 times faster than IBM's. And half the size.

Take two personal computers: Hewlett-Packard's new Series 200 Model 16 and the IBM PC. Run the same software—Context/MBA™—and solve the same problem.

You'll see one big difference. It takes the IBM minutes to do what the HP 200 can do in seconds* And that's just the beginning.

With Context/MBA™ integrated software, the HP 200 can do word processing, spreadsheet projections, indexing, filing and graphics all at one time. Change one number in a projection... and the graphs change, automatically. It's that easy. And with the HP 200, it's just about that fast.

Because, with its super-powerful MC68000 micro-processor, the HP 200 can address 32 pieces of information at one time. (Twice as much as the IBM.) And it's got the memory to handle big, complicated problems. (You can have as much as 768Kb.)

But for all its power, the HP 200 is surprisingly small. The whole computer, not including the keyboard, measures just 12″ across by 13″ deep. So, putting a personal computer on your desk doesn't mean taking everything else off.

If you'd like to see how the HP 200 measures up, call (800) 547-3400. (In Alaska, Hawaii and Oregon call (503) 758-1010 or TTY (503) 758-5566.) Ask operator #411 for more information and the name of the HP dealer or sales office nearest you.

HEWLETT PACKARD

TC-02310

*Recalculating and refreshing ■ screen with words, numbers and graphics. Both ran equivalent Context/MBA™ programs.

Context/MBA™ is a trademark of Context Management Inc. IBM is ■ trademark of International Business Machines.

tells of it. "If there is such a charged particle, it should become visible in the Wilson cloud chamber, should it not?" Kibuchi asks. "I answered, 'Yes, the particle can be found in the cosmic rays.'"

There the pi meson was duly found, after a few years of unraveling some puzzles. Today the simple Yukawa equation has given way to a grand scenario of particles found and conjectured in entire families, much extending Yukawa's pioneer path. His penetrating self-analysis is nicely translated. L. M. Brown of Northwestern has given a knowing introduction to Japanese physics in the times of the narrative, an account that makes it quite clear that Yukawa, although he was a self-contained worker, was by no means alone. The book is published now in English by a new Singapore firm, determined to make available good books (most of them original advanced treatises) at low prices for the students, many in Asia, from whose work will flow the worldwide physics of the future.

A book so personal may justify the self-indulgence of the physicist-reviewer, who first met Hideki and Sumi Yukawa in that year of grim resolution 1945 at their Kyoto house, over a ceremonial cup of tea. The gauche young visitor, in the uniform of the occupier, came half-triumphant, half-guilty, to receive a graceful welcome at their hands. They gave what they had in war-exhausted Japan, quiet, beauty and the celebration of peace, and it was treasure.

BIOMINERALIZATION AND BIOLOGICAL METAL ACCUMULATION: BIOLOGICAL AND GEOLOGICAL PERSPECTIVES, edited by P. Westbroek and E. W. De Jong. D. Reidel Publishing Company. Distributed in the U.S. and Canada by Kluwer Boston, Inc., 190 Old Derby Street, Hingham, Mass. 02043 ($69.50). Hard-shelled or spiny, or sharp-toothed and stout of bone, many a member of the animal kingdom has manifested control over and investment in the austere kingdom of minerals. Add examples such as eggshell and the pathological kidney stone, to say nothing of the tiny otoliths within the orienting instrumentation of animals from sharks to human beings, and the case is stronger. About half of the space in this varied and prompt report of an international congress held in the Netherlands only a year ago is given over to papers touching many aspects of calcification. Plainly the examples just listed are signs of the liminess of the present world of life, organisms adept at mobilizing the ions of calcium up to the crystalline level.

The topics are not new; indeed, this was the fourth interdisciplinary congress of a series, a gathering of biologists, geologists and chemists. This time their brief was widened. A third of the papers treat the accumulation of minerals other than those of calcium, and the rest of the 500-page volume, its attitude closer to plate tectonics than to electron micrographs and polysaccharide chemistry, inquires into the global implications of the ancient alliance between minerals and organisms.

There are now known about 40 minerals normally formed by extant organisms, plus nearly 20 others that appear as pathologies (including dolomite, a bladder stone of the Dalmatian dog, whose spots are not its only claim to oddity). Our bone crystallites are mainly a mineral of calcium phosphate, hydroxyapatite; the shelly stuff of mollusks is calcium carbonate, in one or another of its crystal forms; also on the list are pyrite, galena, fluorite, barite, gypsum and magnetite. It is true that some of these relations may be less than intimate: the organism induces the growth of the mineral in a pseudoinorganic manner, so to speak on the side.

Even more commonly crystal growth is more or less controlled by a little-understood assembly of specific organic macropolymers, which act to determine the mineral phase, its crystallographic orientation and its overall form. Tiny opal-like silica spheres somehow are assembled into the specific and lovely forms of the diatoms and the radiolarians. An organic framework can be seen to precede the opaline deposit and may act as a template. Life is good at constructing surface layers, a strategic device for the control of crystal growth, and there are reports here of the strong inhibition of crystal growth, even in vitro, by specific organic products that disturb nucleation. Electron micrographs at almost the resolution of the atomic layers disclose that in the tiny calcite particles of the inner ear of a rat the crystallites are oriented to yield a single-crystal diffraction pattern, even though actually the structure is an ordered overlap of several tiny crystals. The full terms of the alliance between the two species of order remain undisclosed.

The tidal flats of our times are not much different from the world of the early Precambrian, before life could leave the wet sea edge. Nowadays the evolved competitors are keen, but even so some colonial blue-green algae are found on salty shores worldwide from the Bahamas to western Australia. One remarkable study examined these matted structures over time; the living forms lithify to dominate the intertidal surfaces. These crusts of calcium carbonate form through an elaborate process of seasonal change; they resemble in detail the stromatolites of a billion or two years ago and more. The process that preserves the structure destroys the very organisms that built it; the alliance was intimate even then, although not yet easy.

The sovereign cycles of the earth are plainly two: at depth gravitational and radioactive heat drives a slow, majestic flow, to raise mountains, open and close seas and outgas deep-lying materials within the huge volume of the planet. On the surface there is a thin shell of life, sun-nourished, catalytically remarkably active, demonstrably evolving in complexity and perhaps in gross activity over the entire span of planetary history. One molecule, carbon dioxide, plays two roles. It is the nutrient that feeds the photosynthetic elaboration of the carbon skein of life. It is also the single atmospheric component most important in the regulation of the temperature of the earth's surface by its greenhouse control over outgoing infrared.

Life demands enough of carbon dioxide, and yet life cannot easily survive too much of it. Most of the carbon atoms now near the surface of the earth are in fact buried deep in the calcite jetsam of the waters of all geologic time. Otherwise our planet might be an oven-hot second Venus. For four billion years the right balance has somehow been struck. Was it a subtle feedback loop set up within the fabric of life itself? Or was the balance maintained by blind geology? The terrestrial thermostat has maintained a living planet even though the sun has slowly grown more luminous, on its way to the swollen red giant of the future.

That might be reflected in the uptake of carbon dioxide by calcium during the weathering of the exposed feldspars in crystalline rock. Hot climates mean faster weathering, less free carbon dioxide and a cooling climate, a feedback loop entirely independent of life. Such an earth cannot freeze over, either; the gas from the depths is fed out even through a glaciated surface, but weathering would not persist under an ice blanket. Or is it that weathering rates depend heavily on the transport of carbon dioxide to the rocky surfaces by the vigorous life in the soil of a hot and sunny climate? Evolution or good fortune? Gaia or Pluto?

This big book holds at least half a dozen readable review papers over the entire set of issues mentioned. Another dozen or so papers are straightforward enough to instruct and even amaze the general reader. The remaining score, terse and technical, can reward the browser. One report by a young Boston University biologist indicates by direct simulation with living cyanobacteria (a sample she took from the Charles River) that the very old gold ores of the Witwatersrand may indeed have been formed through direct microbial involvement. Is glittering gold another biomineral of the microbial mats?

Envisioning automobiles that will slip through the wind is only a beginning. The men and women who work the assembly lines then translate those visions into reality. That's why, in 68 plants across the country, over 1100 Employee Involvement Groups have been established by Ford Motor Company and the UAW. Their goal? To build the highest quality cars and trucks in the world. Members of this group, at the Kansas City, Missouri, Assembly Plant, used their skills to refine the fit and finish procedures needed to build the brand-new Ford Tempo well before this highly aerodynamic car went into production. All part of the effort to get a sound body designed to be free of squeaks and rattles. Employees speaking and management listening. Another reason Ford Motor Company achieved the highest quality rating of any major American car or truck maker. A fact based on a national survey measuring things gone wrong, after 3 months of ownership, conducted among 11,000 people owning Ford and competitive 1982 new vehicles. And it's further evidence that Ford Motor Company employees are great body builders.

Body Builders.

Ford
Mercury
Lincoln
Ford Trucks
Ford Tractors

Quality is Job 1.

Ford

Established 1845 SCIENTIFIC AMERICAN June 1983 Volume 248 Number 6

Physical Disability and Public Policy

The civil rights of disabled Americans require a more accessible environment. The present Administration, however, hesitates to enforce laws calling for the removal of architectural barriers

by Gerben DeJong and Raymond Lifchez

In the past 15 years or so there has been a major change in the policy of the U.S. with respect to citizens with disabilities. Between 1968 and 1980 Congress passed a series of laws aimed at enhancing the quality of life for disabled Americans through various measures, including the removal of architectural barriers, the improvement of educational opportunities for disabled children and the encouragement of self-help organizations for people with disabilities. The centerpiece of the new legislation was the Rehabilitation Act of 1973. Section 504 of the act, which bars discrimination on the basis of handicap in any program or activity benefiting either directly or indirectly from Federal financial aid, has emerged as the most controversial and most litigated piece of the legislation. In essence the Rehabilitation Act extends to disabled Americans a previously uncodified form of civil-rights protection. Accordingly the 1973 law is sometimes called the Civil Rights Act of the Handicapped.

In spite of the clear intent of Congress in the matter the present status of disabled people in American society is uncertain, in large part because the Reagan Administration has been reluctant to enforce the new laws. Meanwhile there are signs of a public backlash: after an initial period of little opposition to these measures the nation seems to have entered a period of doubt, particularly over the cost of making the built environment more accessible to people with disabilities. At a time of widespread retrenchment in Government expenditures for social programs Americans with disabilities are finding themselves increasingly on the defensive.

In what follows we shall review the background of this issue, summarize what is known about the dimensions of disability in the U.S., look into the causes of the current stalemate in the implementation of the existing disability laws and try to identify some of the ingredients needed to make the nation's disability policy a viable one in the face of current political and economic realities. In particular we shall focus on one highly visible aspect of the problem, namely the removal of architectural barriers to the progress and well-being of people with disabilities.

The disability legislation of the past 15 years is based on familiar civil-rights principles such as equal opportunity, nondiscrimination, integration, free choice and self-help, to name a few. The thread linking all these principles is the concept of access—access to education, to employment, to public facilities and services, to transportation, to housing and to other resources needed by disabled people to more fully realize their rights as citizens. Architectural barriers compromise access. They are seen in this context as obstacles that limit opportunity, promote discrimination, prevent integration, restrict choice and frustrate self-help.

In a sense the new disability legislation also has an affinity with the traditional noninterventionist values of the political right. The legislation emphasizes self-reliance and the right of disabled individuals to shape their own destiny. The aim of the legislation is merely to remove the barriers that prevent disabled people from achieving that goal. Although the legislation does in fact impose new burdens on certain segments of American society, the "bootstrap" philosophy implicit in the laws is unmistakable.

Much of what is known about disability in the U.S. comes from the Health Interview Survey, a continuing nationwide study sponsored by the National Center for Health Statistics. Each year the center, working in collaboration with the Bureau of the Census, selects a sample of about 40,000 households, representing approximately 120,000 persons. One respondent from each household is interviewed in depth to ascertain the state of health of each resident, including any functional limitations attributable to a chronic health condition. According to the Health Interview Survey for 1979 (the latest year for which fully tabulated results are available), some 31.5 million Americans, or 14.6 percent of the noninstitutionalized population, are limited in some way by a chronic health condition. An estimated 7.9 million (or 3.7 percent) are considered severely disabled, that is, they are unable to carry on some major activity such as attending school, working or housekeeping. As one might expect, disability increases significantly with age: 46 percent of those who are 65

OUTSTANDING EXAMPLE of a public place designed to be accessible to people with disabilities is seen in the photograph on the opposite page. The view is of a rooftop "vertical park" adjacent to the new Provincial Government Offices and Law Courts Building in Vancouver, B.C. The stairway, which provides access to the terraced, low-rise building at several levels, incorporates a series of diagonal ramps suitable for use by people in wheelchairs. The ramps have a grade of 8 percent (that is, they rise eight feet for every 100 feet of their length), in compliance with the local building code. The $100 million civic center was designed by Arthur Erickson Architects of Toronto. The client was the British Columbia Building Corporation.

years old or older have a chronic activity-limiting disability, and 16.9 percent are unable to carry on a major activity.

For more detailed statistics on the prevalence of disability in the U.S. one must turn to another source: the Survey of Income and Education, which was conducted in 1976 by the Census Bureau. The results of this survey are based on interviews with people representing 158,500 households. For the purpose of the survey people were considered disabled if they had a chronic health condition that prevented them from participating in a major activity appropriate to their age group.

The Survey of Income and Education indicates that blacks are significantly more likely to be disabled than whites: 17.6 percent of the black population surveyed were found to be disabled, compared with 13.7 percent of the white population. A more striking contrast can be found in the 18-to-64 age group, in which 19.4 percent of the black population and 12.6 percent of the white population were found to be disabled. People classified as Hispanic were found to have the lowest rate of disability: 10.6 percent.

According to the same survey, the prevalence of disability varies inversely with education. In the 18-to-64 age group 38.5 percent of those with less than an eighth-grade education, 10.3 percent of those with a high school education and 5.3 percent of those with a college degree were disabled. Because the ability to work is central to the definition of disability for this age group, a person's level of education can significantly affect his or her ability to obtain work with or without a long-term health condition.

Adopting the Federal poverty line as a benchmark, the Survey of Income and Education found that poor people in the 18-to-64 age group were much more likely to be disabled than nonpoor working-age people (28.7 percent as against 11.8 percent). Another way of stating the relation between disability and poverty for this age group is to point out that disabled people were more than twice as likely to be poor than nondisabled people (18.7 percent as against 7.1 percent).

The direction of causation between poverty and disability is not well understood. Does poverty cause disability or does disability cause poverty? Based on national survey data from the 1960's Howard Luft of Stanford University estimates that at least 30 percent of the Americans with disabilities are poor because of their health problems. He reports that among white males the figure rises to 75 percent. Luft argues that the causal sequence from poverty to disability is less clear and cannot be easily disentangled from the effects of other factors such as age, sex, race and education.

The prevalence of disability varies by as much as 45 percent from one region of the country to another. Data from the Survey of Income and Education indicate that for people three years old and older New England has the lowest rate of disability (12.6 percent), whereas the East South Central region has the highest (18.3 percent). Among the contiguous 48 states Nebraska has the lowest disability rate (11.1 percent); West Virginia has the highest (21.9 percent).

Most of these statistics reveal little about the demands that disability places on the nation's system of health and human services. The two largest Federal programs for disabled people in the U.S. are the Social Security Disability Insurance program and the Supplemental Security Income program, both of which are managed by the Social Security Administration. In 1980 these two income-transfer programs made

YEAR	PUBLIC LAW NO.	TITLE OF LAW	KEY PROVISIONS
1968	90–480	ARCHITECTURAL BARRIERS ACT	Requires that buildings built with Federal funds or leased by the Federal Government be made accessible
1970	91–453	URBAN MASS TRANSPORTATION ACT	Requires eligible local jurisdictions to plan and design accessible mass-transportation facilities and services
1973	93–87	FEDERAL AID HIGHWAY ACT	Requires that transportation facilities receiving Federal assistance under the act be made accessible; allows highway funds to be used to make pedestrian crosswalks accessible
1973	93–112	REHABILITATION ACT	Prohibits discrimination against qualified handicapped people in programs, services and benefits that are Federally funded; creates Architectural and Transportation Barriers Compliance Board
1975	93–391	DEPARTMENT OF TRANSPORTATION APPROPRIATIONS ACT	Prohibits purchase of mass-transit equipment or construction of facilities unless they are accessible to elderly and handicapped people
1975	94–103	DEVELOPMENTAL DISABILITIES ASSISTANCE AND BILL OF RIGHTS ACT	Establishes protection and advocacy systems for developmentally disabled people; establishes representative councils in each state for developmentally disabled people
1975	94–142	EDUCATION FOR ALL HANDICAPPED CHILDREN ACT	Provides for a free appropriate education for handicapped children in the least restrictive setting possible
1975	94–173	NATIONAL HOUSING ACT AMENDMENTS	Provides for the removal of barriers in Federally supported housing; establishes Office of Independent Living for disabled people in U.S. Department of Housing and Urban Development
1978	95–602	REHABILITATION COMPREHENSIVE SERVICES AND DEVELOPMENTAL DISABILITY AMENDMENTS	Establishes independent living as a priority for state vocational rehabilitation programs; provides Federal funding for independent-living centers
1980	96–265	SOCIAL SECURITY DISABILITY AMENDMENTS	Removes certain disincentives to work by allowing disabled people to deduct independent-living expenses in computing income benefits

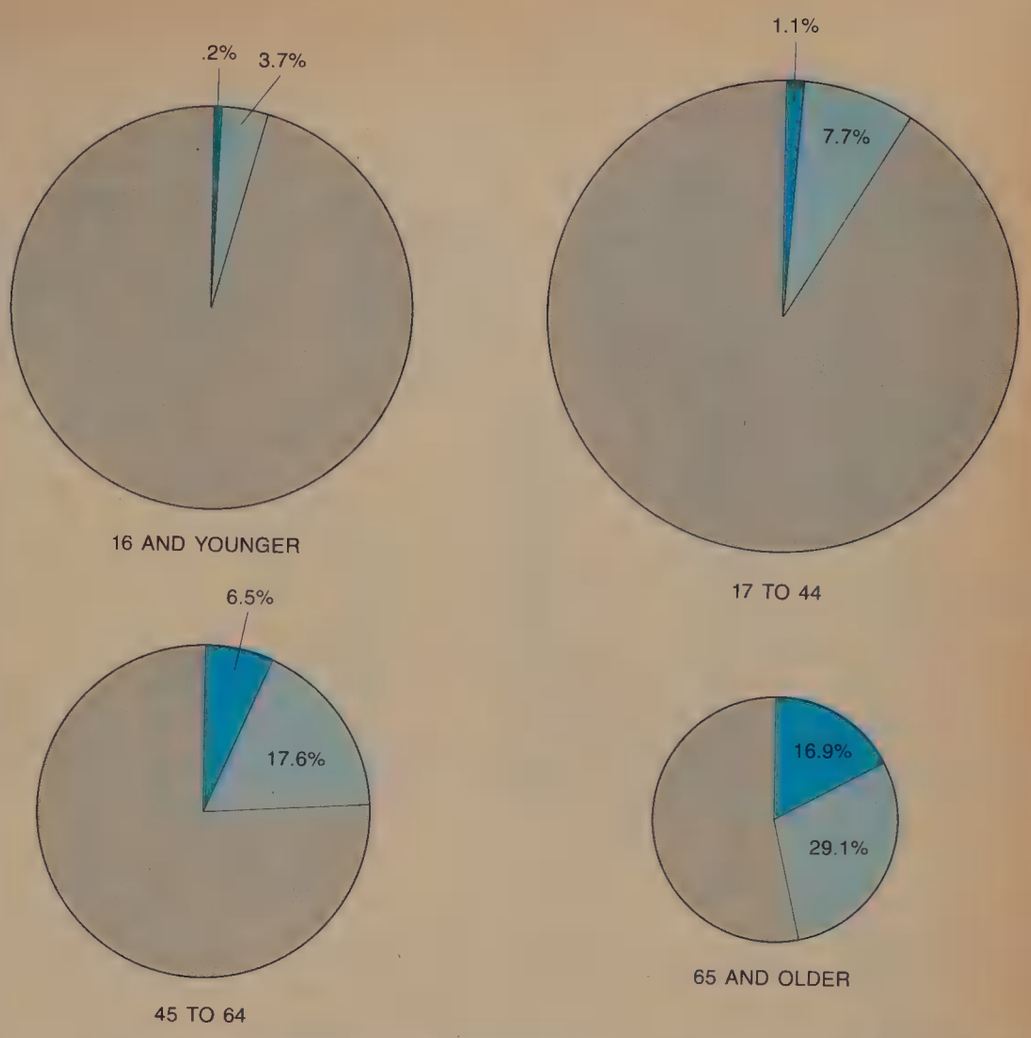
FEDERAL DISABILITY LAWS passed by Congress between 1968 and 1980 marked a shift in the public policy of the U.S. from one that focused primarily on providing personal-adjustment services to disabled individuals to one that concentrated more on making the built environment accessible to all Americans. In addition to the major pieces of legislation listed here Congress enacted numerous other measures aimed toward the same objective. Much of this legislation has not been fully implemented by the Reagan Administration.

$20.6 billion in cash payments to more than four million working-age people with disabilities.

Monroe Berkowitz of Rutgers University estimates that disability-related expenditures of $63.5 billion were made on behalf of working-age people in 1977, the latest year for which aggregate cost data are available. Of this figure $47.6 billion went to public and private income-transfer payments, $12.9 billion went to medical care and $3 billion went to direct services and various labor-market interventions such as job programs and vocational rehabilitation. None of these figures includes transfer payments or services on behalf of children or people of retirement age.

The number of people who receive disability-related income-transfer payments has stabilized in recent years. Nevertheless, the problem of chronic disability continues to have a disproportionate and growing impact on the health-care system. Data from the 1979 Health Interview Survey indicate that although people with a chronic activity limitation constitute less than 15 percent of the population, they account for 29.3 percent of all visits to physicians, 40.1 percent of discharges from short-stay hospitals and 58 percent of days spent in short-stay hospitals. These percentages all represent increases from previous years.

In spite of the prominence of the issue of environmental accessibility in recent discussions of disability policy, little is known about the size of the disabled population for whom environmental accommodations are required. An important statistic to consider in this context is the number of "mobility-impaired" people in the population.



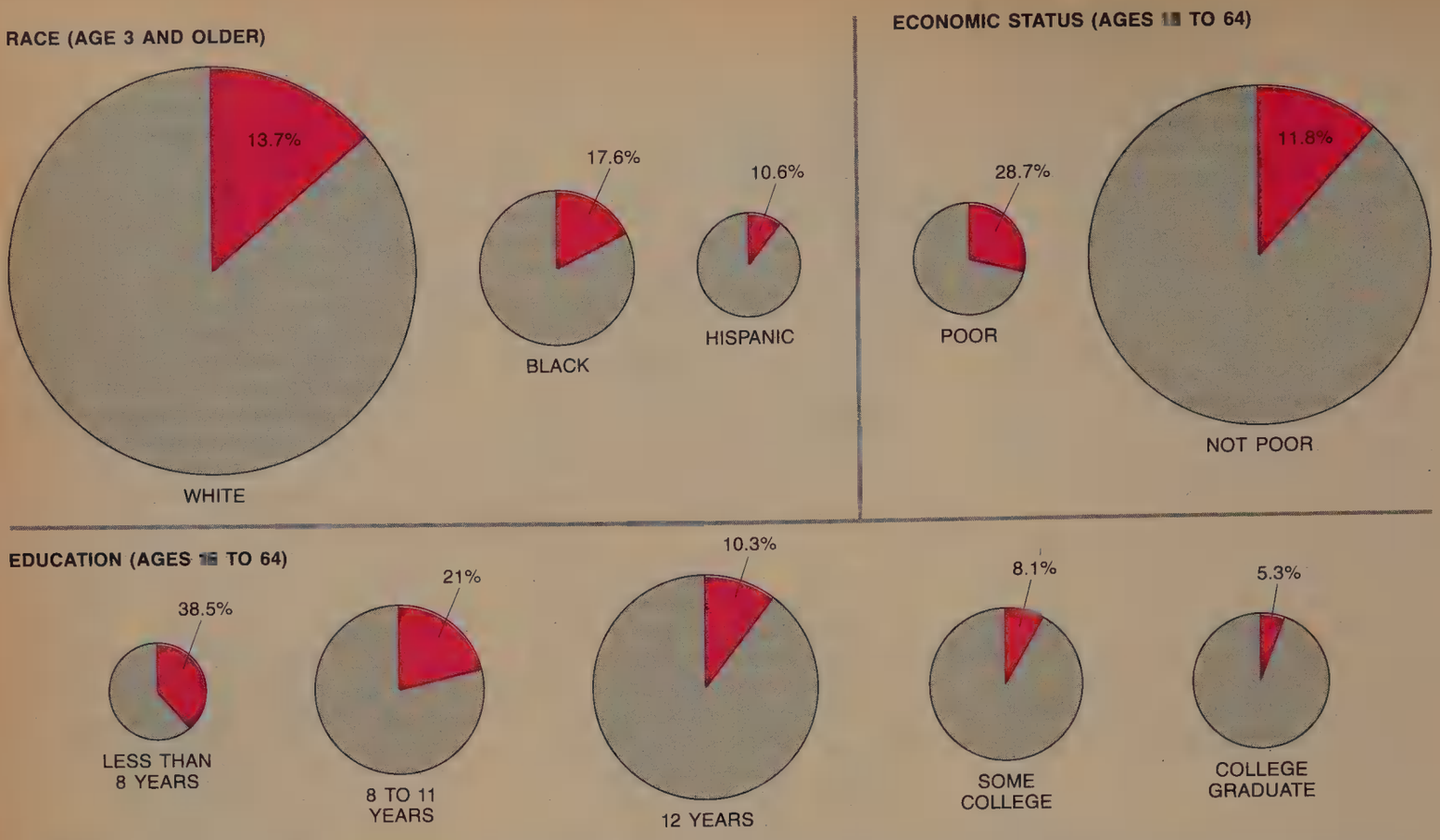
PREVALENCE OF DISABILITY increases significantly with age. The area of each of the circles in this chart is proportional to the number of Americans in the associated age group. The colored sections show the relative size of the disabled population in each age group. The light-colored sections correspond to the percentage of people with a limited disability, the dark-colored sections to the percentage of people with a severe disability. The chart is based on the 1979 Health Interview Survey, conducted by the National Center for Health Statistics. For this purpose people were classified as severely disabled if they were unable to carry on a major activity appropriate to their age group, such as attending school, working or housekeeping.

One indicator of mobility impairment is the employment of various devices to aid mobility. According to the 1977 Health Interview Survey, about five million people in the noninstitutionalized civilian population rely on one or more devices of this kind. Another million or so people in nursing homes use mobility aids or are confined to bed. It is fair to say that the problem of physical accessibility is most acute for these six million people. In addition there are thousands of Americans who do not use mobility aids but are nonetheless limited in their mobility because of a chronic health condition. A common example would be the person with coronary heart disease who cannot climb stairs without inducing a painful attack of angina.

Most striking, we believe, is the recent growth in the use of mobility aids. Comparative data from the 1969 and the 1977 Health Interview Survey suggest that there has been a 40 percent growth in the use of such aids in this period. The same source indicates that there are some 645,000 people in the noninstitutionalized civilian population who use wheelchairs. Based on our knowledge of selected disabling conditions, we believe the results reported by the Health Interview Survey are somewhat low. Our impression is supported by the Social Security Administration's 1978 Survey of Disability and Work among people in the 18-to-64 age group. The Social Security Administration reports that there are 907,000 people of working age alone who use wheelchairs. No estimate is given for children or older people. As for the institutionalized population, the National Center for Health Statistics, based on its 1977 National Nursing Home Survey, reports that there are 517,000 people in nursing homes who use wheelchairs.

Most surveys of health and disability provide only an instantaneous view of disability in the U.S. More interesting from the standpoint of disability policy is the fact that the prevalence and composition of disability has changed over time. One reason for the change is the shifting age structure of the population. Since the turn of the century there has been a 27-year increase in life expectancy at birth: from 47 years in 1900 to 74 years in 1980. As a result there has been an eightfold increase in the size of the population 65 years old and older: from three million (representing 3 percent of the population) in 1900 to 25 million (representing 11 percent of the population) in 1980. As the life span has increased, so has the probability that an individual will acquire a major disability-inducing disease, and as the population of the elderly has increased, so has the prevalence of disease and disability.

The 1930's marked a major turning point in the medical management of chronic disease and disability. Until then people seldom died because of the disease or disability; usually they died because of secondary infections such as pneumonia or urinary-tract disease. The introduction of sulfonamide drugs in the late 1930's and penicillin and other antibiotics in the 1940's mitigated such terminal infections. The net effect was that the prevalence of certain diseases and disabilities increased because their average duration was prolonged.



DEMOGRAPHIC ASPECTS of disability in the U.S. are represented here in terms of race, economic status and education. The statistics, which are derived from the Survey of Income and Education, conducted in 1976 by the Bureau of the Census, show that people who are black, poor or less educated are more likely to be disabled than the rest of the population. Again the area of each circle is proportional to the number of Americans in each category; the colored section in this case represents the total percentage of disabled people.

Although it has now been more than four decades since the introduction of sulfonamides, health surveys still reveal large increases in the prevalence of disability. According to Alain Colvez and Madeleine Blanchett of the Quebec Ministry of Social Affairs, the results of the Health Interview Survey indicate that the number of Americans with a chronic activity limitation increased by 8.2 million, or 37.3 percent, from 1966 to 1976; meanwhile the general population increased by only 10 percent.

More impressive was the large increase in severe disability. Colvez and Blanchett report that the number of Americans unable to carry on their major activity because of a chronic health condition increased by 3.4 million, or 83.2 percent, during the 10-year period. In terms of prevalence rates, the rate of severe disability increased from 213 people per 10,000 in 1966 to 355 per 10,000 in 1976: an increase of 66.7 percent. More recent data from the Health Interview Survey indicate that by 1979 the prevalence rate had increased to 365 people per 10,000: an increase of 71.4 percent since 1966. The leading causes of activity restriction have also changed over this period. Arthritis and rheumatism have replaced heart disease as the leading cause of disability, and diabetes has become the fastest-growing disabling condition among the leading causes of disability.

Colvez and Blanchett argue that none of these developments can be attributed solely to factors such as methodological changes in the Health Interview Survey, changes in the respondent's perception of what constituted disability or changes in the size or overall age distribution of the U.S. population. Although we cannot offer a comprehensive reason for the findings reported by Colvez and Blanchett, we believe at least some of them can be attributed to the introduction of new medical interventions that are now saving lives at the price of longer-term and severer residual disabilities. In addition new interventions are in some cases significantly extending the lives of those already disabled.

The combination of these two trends augurs a variety of new demands facing American society in the form of new technology and highly intensive services: nonvocal communication devices, respirator management, 24-hour attendant care and a variety of community-based long-term care services. At present few of these services are routinely available; usually they can only be acquired with great difficulty from multiple providers and funding sources.

These developments present important ethical and public-policy issues that cannot be ignored in any consideration of the demands of disabled people for fuller access to society and its institutions. Many questions have already been asked about the quality of life that can be expected for those whose existence appears marginal. These questions, however, implicitly place the onus on the disabled individual rather than on society, whose services can in many cases materially affect the quality of life. What about society's obligation to provide a minimum quality of life for those who survive as the result of medical research and technological advances that have often been publicly demanded, publicly supported and publicly financed? The financial consequences of these public decisions are so overwhelming that few individuals or families are capable of assuming the financial burden for services and environmental supports needed to provide a minimum quality of life.

In spite of the growing number of people with severe physical disabilities, the total number remains small compared with the rest of the population. Moreover, the needs of such people are varied and often highly specialized, making it difficult to form an effective political constituency representing these needs. Meanwhile this segment of the population depends on the limited awareness of a larger public, which is hesitant to provide needed services and accommodations. It is not widely understood that these needs arise because society has been successful—both in saving and in

prolonging lives. The new social needs are the price of that success.

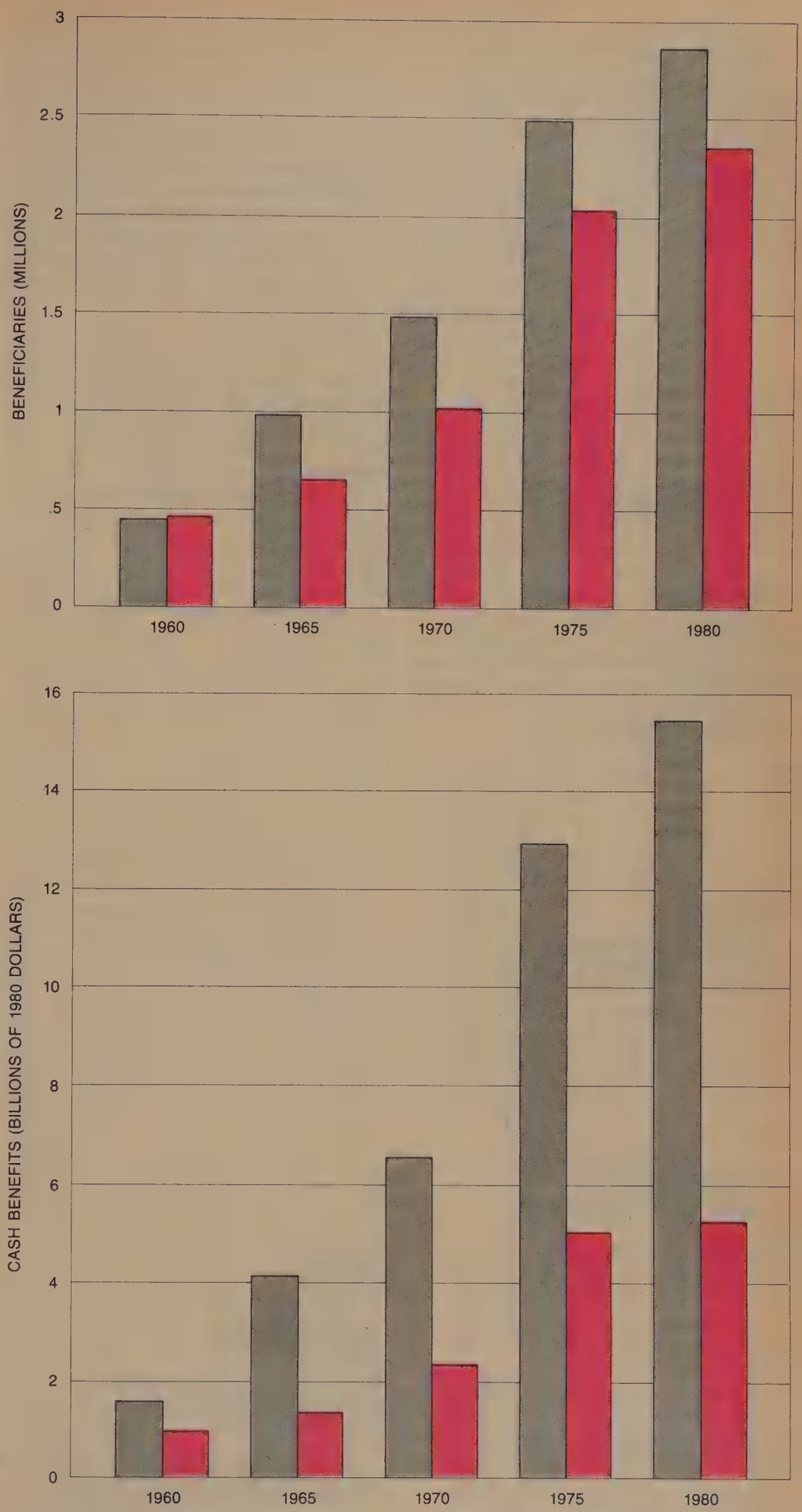
The disabled population in the U.S. has also undergone a political and social transformation in the form of a social movement commonly referred to as the independent-living movement. This movement has changed how disabled people view themselves and how they wish to be viewed in American society: no longer as passive victims deserving of charitable intervention but as self-directed individuals seeking to remove environmental barriers that preclude their full participation in society.

The independent-living movement had its origins in the early 1970's when small groups of disabled people banded together to form self-help organizations known as independent-living centers. The first centers were established in Berkeley, Calif., and Boston, Mass. Today about 100 independent-living centers across the country provide or arrange for a variety of services such as housing referral, transportation, attendant care, peer support and legal advocacy. The leaders of the independent-living movement have also been active on the political front: in promoting the passage of disability legislation, in seeking to make local transit systems more accessible, in removing architectural barriers and in changing the awareness of political leaders and other decision makers about the particular needs of disabled people.

The independent-living movement offers a radically different view of the problem of disability and its solution. According to representatives of the movement, the problem of disability is one not only of physical impairment but also of unnecessary dependence on relatives and professionals, of architectural barriers and of unprotected rights. In this view the pathology is not in the individual, as the medical model would suggest, but rather in the physical, social, political and economic environment that has up to now limited the choices available to people with disabilities. The solution to these problems is not more professional intervention but more self-help initiatives leading to the removal of barriers and to the full participation of disabled people in society.

The independent-living movement should not be viewed as an iconoclastic one but rather as one that affirms traditional American values such as self-reliance. The movement's commitment to the nation's political and economic institutions is remarkable, considering how disabled people have been excluded from the mainstream of the nation's political and economic life. Disabled people still want to be part of the system. They want access.

The barrier-free movement—a shorthand term to describe the national effort



TWO LARGEST FEDERAL INCOME-TRANSFER PROGRAMS for disabled people of working age have grown rapidly over the past two decades, as this pair of charts shows. The two programs represented are the Social Security Disability Insurance program (*gray bars*) and the Supplemental Security Income program and its predecessors (*colored bars*). Both programs are now run by the Social Security Administration. Between 1970 and 1980 the number of disabled people receiving benefits under one or both of the programs doubled (*top chart*). Since 1977 the number of beneficiaries in both programs has leveled off. The total benefits paid under both programs also rose sharply, reaching a total of $20.6 billion in 1980 (*bottom chart*).

to eliminate architectural barriers—actually preceded the rise of the independent-living movement. Although the exact moment of conception cannot be determined, the barrier-free movement goes back at least to 1958. That was the year the President's Committee on Employment of the Handicapped, with the help of the Veterans Administration, published a tentative guide to help make public buildings accessible to people with impaired mobility. The effort, which depended entirely on voluntary action, produced no discernible results.

A milestone year for the barrier-free movement was 1961. That was the year the American National Standards Institute (A.N.S.I.), a private architectural standard-setting organization, issued Standard A117.1. The standard specified the minimum requirements and working details for structures such as walkways, parking spaces, ramps, stairs, floor surfaces, mirrors, water fountains, public telephones, control identifications and warning signals. Work on the A.N.S.I. standard was sponsored by the President's Committee on Employment of the Handicapped and by the National Easter Seal Society for Crippled Children and Adults.

The promulgation of the A.N.S.I. standard in 1961 and the widespread endorsement it received generated interest in the architectural-barriers problem and created the expectation that the problem would be speedily addressed. On the Federal level, however, no action beyond voluntary compliance was immediately taken to adopt the standard with respect to Federal construction. In response, as part of the 1965 Amendments to the Vocational Rehabilitation Act, Congress established the National Commission on Architectural Barriers to Rehabilitation of the Handicapped in order to study the problem.

The commission's study and three concurrent surveys found compliance to be minimal. Buildings constructed and renovated with Federal funds remained inaccessible. At the local level only 3 percent of the cities surveyed had adopted accessibility codes, and roughly two-thirds of the architects surveyed were unfamiliar with accessibility standards. More than anything, these surveys demonstrated that voluntary compliance alone would not yield the intended results.

The shortcomings of voluntary compliance led Congress to pass the Architectural Barriers Act of 1968. The act's stated purpose was to "ensure that certain public buildings, financed with Federal funds, are so designed and constructed as to be accessible to the physically handicapped." The act delegated authority for setting accessibility standards to various Federal departments and agencies. There was no attempt to add to or take away from what had been defined as the minimum standards of access already accepted in principle by Federal agencies since 1961. In fact, each of the delegated Federal agencies promptly adopted the 1961 A.N.S.I. standard as the standard for their respective regulations.

The task of enforcing the 1968 act, however, remained unresolved until the enactment of the Rehabilitation Act of 1973. Section 502 of the Rehabilitation Act established the Architectural and Transportation Barriers Compliance Board, which was authorized to investigate complaints, hold public hearings, issue compliance orders and seek enforcement of its orders by the courts if necessary.

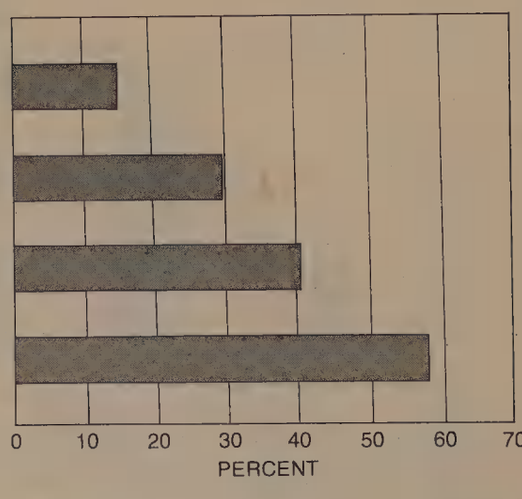
The legal machinery to address the problem of architectural barriers was strengthened in 1977 when regulations pursuant to Section 504 were issued by the Department of Health, Education, and Welfare. The regulations implementing Section 504 required that programs receiving Federal financial assistance be made accessible to, and usable by, disabled people in order to provide them with effective services. The regulations specified that new facilities constructed after June 3, 1977, would have to meet accessibility standards and that existing facilities would have to be made accessible by June 2, 1980. The regulations also specified that conformance with the A.N.S.I. standard would constitute compliance, but other methods would be allowed if they provided equivalent access. The effect of the regulations was broad, since Federal financial assistance extends to many areas such as education, transportation facilities, public housing and other activities affecting the lives of disabled people.

The Federal lead with respect to architectural accessibility has been followed by the states. All 50 states have enacted laws requiring that certain facilities be made physically accessible to disabled people. According to information compiled by the National Center for a Barrier Free Environment, approximately half of the states rely either on the original 1961 A.N.S.I. standard or on a slightly modified 1971 version. The other half have adopted their own standards or have borrowed standards from another state.

A survey of state accessibility laws in 1979 by the American Bar Association found that state legislation is not equally forceful throughout the country. The survey made a distinction between two types of state accessibility laws. Laws of the first type are classified as declaratory: they merely assert that disabled people have a right of access. According to the bar association, such laws "have consistently been held by the courts to be insufficient manifestations of legislative intent to require any environmental modifications to facilitate such use." Laws of the second type are considered more affirmative: they "vary from simple delegations of authority to promulgate rules, to lengthy acts incorporating minimum standards... and specific administrative and enforcement methodologies."

One of the most troublesome problems faced by architects in trying to eliminate architectural barriers in their designs is the application of accessibility standards to different settings. For example, there is little dispute over the specifications for an accessible bathroom, but there is no agreement on how many accessible bathrooms there should be or how they should be spaced throughout a building. It is the implementation and application of standards that have caused major difficulties for practicing architects.

Although most Federal, state and lo-



UTILIZATION OF HEALTH-CARE SERVICES by disabled people is disproportionately high. As these bars for the year 1979 indicate, although people with a chronic activity-limiting health condition totaled less than 15 percent of the U.S. population, they accounted for much higher percentages of all visits to a physician, discharges from short-stay hospitals and days spent in short-stay hospitals. The data are from the National Center for Health Statistics.

cal agencies have embraced the A.N.S.I. standard, many have adopted different criteria for applying it. Hence architects sometimes find themselves faced with conflicting regulations, for example when Federal, state and local agencies join on a public housing project. A similar problem can arise in private commercial construction where the public is served. Even here the architect may be confronted with state and local building codes that have different requirements.

The architect is often caught between the requirements of his client and the demands of people with disabilities. In the past most clients knew or cared little about barrier-free design. Most doubted that there was even a serious demand for accessible facilities. Today there is a much greater awareness of the scope of disability and the need for various types of environmental accessibility.

In spite of increased awareness over the past 15 years, it is not clear that clients are any more interested today in making their facilities accessible. Architects observe that clients often approach the issue of accessibility with one or more of three attitudes: (1) disabled people are not viewed as customers, clients or tenants; (2) barrier-free design is automatically assumed to be too expensive; (3) disabled people are thought likely to stigmatize the facility in the minds of nondisabled users. Few architects are equipped to cope with clients who are so disposed. Meanwhile on the other side of the issue are advocacy groups on behalf of disabled people who are often assumed to be making unreasonable demands.

Unfortunately there are few mechanisms available to reconcile such differences. The problem is not merely the accessibility standards or codes but the failure to provide the technical assistance needed to implement them. By themselves standards and codes are viewed as architectural constraints that invite resistance, if not hostility. In addition there is the constant worry about building inspectors and the constant threat of court action, real or imagined. What is needed is a technical-assistance body that can offer creative solutions meeting both the letter and the spirit of existing standards and codes. What is also needed is a decision-making body that can render these creative solutions and compromises legally binding.

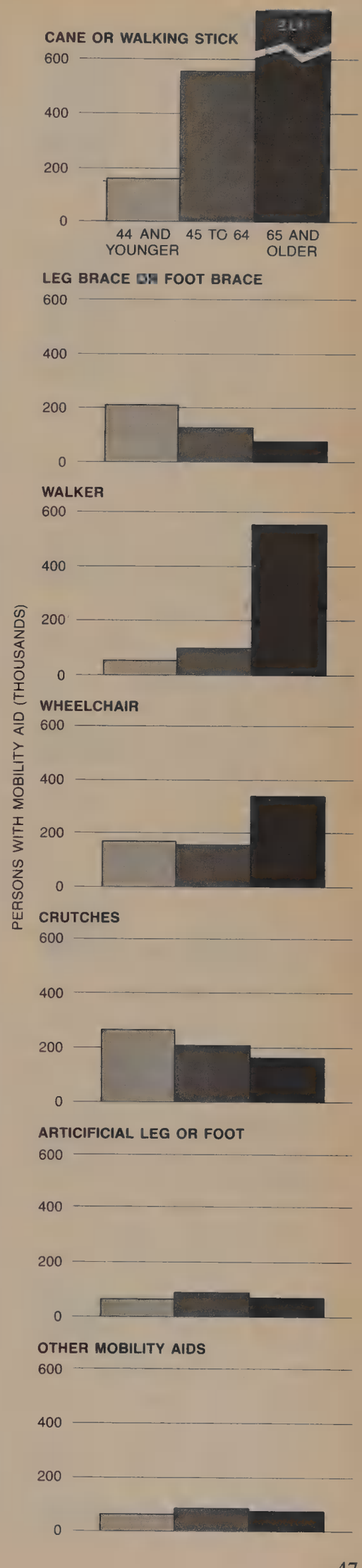
One model of technical assistance is the Special Office for the Handicapped in the North Carolina Department of Insurance. The office includes an architect who works with practicing architects and local building inspectors to administer accessibility standards. As a result the North Carolina architectural-barriers compliance board has received only two requests for waivers over a nine-year period. On the national level the National Center for a Barrier Free Environment, based in Washington, serves as a nongovernmental technical-assistance organization seeking to advise architects, clients and various regulatory boards. Since building codes are largely a state and local matter, however, the greatest need for technical assistance remains at the state and local level.

Architects are not neutral observers merely trying to accommodate the demands of clients and disabled users. Most architects are able-bodied and bring to the design process all the able-bodied attitudes and assumptions that have shaped design concepts in Western culture. Moreover, the architect, like most people, has become accustomed to the social segregation of disabled people in the environment.

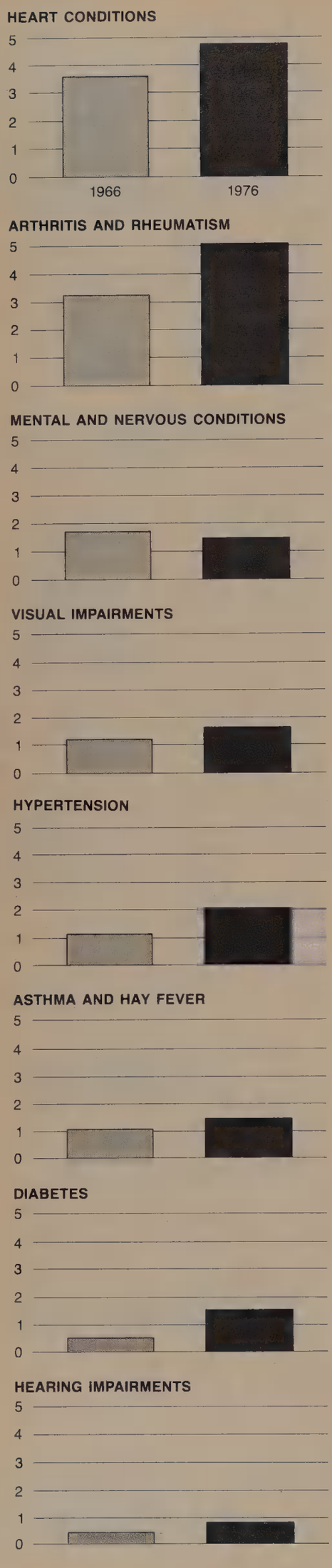
What can be done to correct for this "cultural lag" in professional architecture? We should like to suggest three steps. First, architects must come to know more about disability, just as in recent years they have come to know more about solar energy and climate and how to design with these environmental factors in mind. Architects must come to understand that disabled people are a complex collection of individuals with diverse functional limitations. An important place to start is in the curriculum of both graduate and undergraduate schools of architecture.

Second, architects need to seek guidance from outside their profession. For example, architects are known to enlist the help of behavioral scientists in the design of specialized environments such as housing for the elderly. Now they need to accept the idea that disabled people have something to tell them. Architects need to involve disabled people as design consultants and learn where and how such consultants can be employed.

Third, the architectural profession, through its association and publications, needs to actively support the barrier-free movement and to reward practitioners for outstanding work. So far the profession has viewed the barrier-free movement largely as an infringement on its creative freedom and has treated accessibility standards as a "cookbook" approach to design. The American Institute of Architects, the national professional organization, has never come out openly against accessibility standards, but it has done little to promote the concept of accessibility. Its barrier-free committee was never popular and was



MOBILITY AIDS are now employed by millions of Americans with disabling health conditions, as these bars for the year 1977 indicate. The use of such devices has increased sharply in the U.S. in the past decade or so.



MAJOR CAUSES of disability in the U.S. are represented in this chart for the years 1966 and 1976. In the intervening period the population of the U.S. increased by 10 percent, and the number of people with a disability rose by 37.3 percent. Arthritis and rheumatism have replaced heart disease as the leading cause of disability. Meanwhile diabetes has become the fastest-growing major cause of disability.

eventually disbanded. Even the institute's Bartlett Award, given for excellence in accessibility design, fell from favor because of the careless way winners were selected and because some winning designs did not comply with minimal national standards. Similarly, architectural magazines have done little to promote accessibility as an important design issue. Coverage of barrier-free architecture has mainly been confined to technical or business pages or has appeared in connection with institutional buildings designed for special populations.

Although technical assistance and professional awareness are indispensable, there still is a need at the national level for leadership in developing a comprehensive standard for accessibility. For two decades the 1961 A.N.S.I. standard served as the point of departure for the development of accessibility regulations. The inadequacies of the 1961 A.N.S.I. standard have allowed Federal agencies to apply it in conflicting ways. The price of varying Federal requirements has been high both in dollars and in frustration.

Recognizing the shortcomings of the 1961 A.N.S.I. standard, the Department of Health, Education, and Welfare—in cooperation with the President's Committee on Employment of the Handicapped and the National Easter Seal Society—awarded a contract in 1974 to Syracuse University's School of Architecture to revise and update the 1961 standard. What started as a two-year project became a six-year conflict ending in 1980 when a revised standard was formally adopted by the American National Standards Institute. In spite of the consensus-building efforts of those associated with the Syracuse project, several participants representing Federal agencies voted against the new A.N.S.I. standard on the final ballot.

The failure to achieve a consensus in 1980 set the stage for what has been one of the most acrimonious debates on the issue of environmental accessibility. The General Services Administration and the Department of Health, Education, and Welfare announced the development of a new accessibility standard, as did the Postal Service. In addition the Architectural and Transportation Barriers Compliance Board, as authorized under a 1978 statute, proceeded to develop a set of "minimum guidelines and requirements" that was to serve as the basis for the setting of standards by other Federal agencies. The emerging consensus that had led to the 1980 A.N.S.I. standard suddenly unraveled with the proliferation of conflicting accessibility standards. The situation was further complicated by the arrival of a new Administration that in many ways preferred no regulations at all.

For most of 1981 the conflict was centered on the compliance board. The opening salvo was the board's publication of Minimum Guidelines and Requirements for Accessible Design. The new standard appeared in the January 16 *Federal Register,* just four days before the Reagan Administration took office. The new minimum guidelines were characterized by Vice-President George Bush's Regulatory Review Task Force as overzealous and an example of unnecessary Government regulation. The compliance board soon found itself targeted for extinction by the Office of Management and Budget. The board survived the threat but not without further tumult.

The compliance board was itself divided—11 citizens (nine of whom were disabled) representing the public and 11 members representing various Federal agencies. The scales were tipped when one of the public members was persuaded by the Administration to join the Federal members in a 12-to-10 vote to rescind the minimum requirements. The Federal members argued that the new requirements were too costly. Total recision was averted late in 1981 when compromises were made to eliminate the most costly provisions. The Administration had also come to realize that uniform minimum requirements would simplify rule making for the other Federal standard-setting agencies. On August 4, 1982, the compliance board promulgated its final version of the new minimum requirements.

Although the new guidelines do not go as far as many barrier-free advocates would like, there is at least a consensus on what the minimum requirements should be. The minimum requirements issued by the compliance board are similar to the 1980 A.N.S.I. standard and differ mainly with regard to "scoping provisions," or application criteria, an area largely unaddressed by the 1980 A.N.S.I. standard. Even under the board's guidelines most issues of this kind are left to the Federal standard-setting agencies. Among these agencies only the Postal Service appears to be recalcitrant, as was demonstrated by its vaguely worded proposed regulations published in October, 1982.

In spite of the Reagan Administration's attempt to put the brakes on the

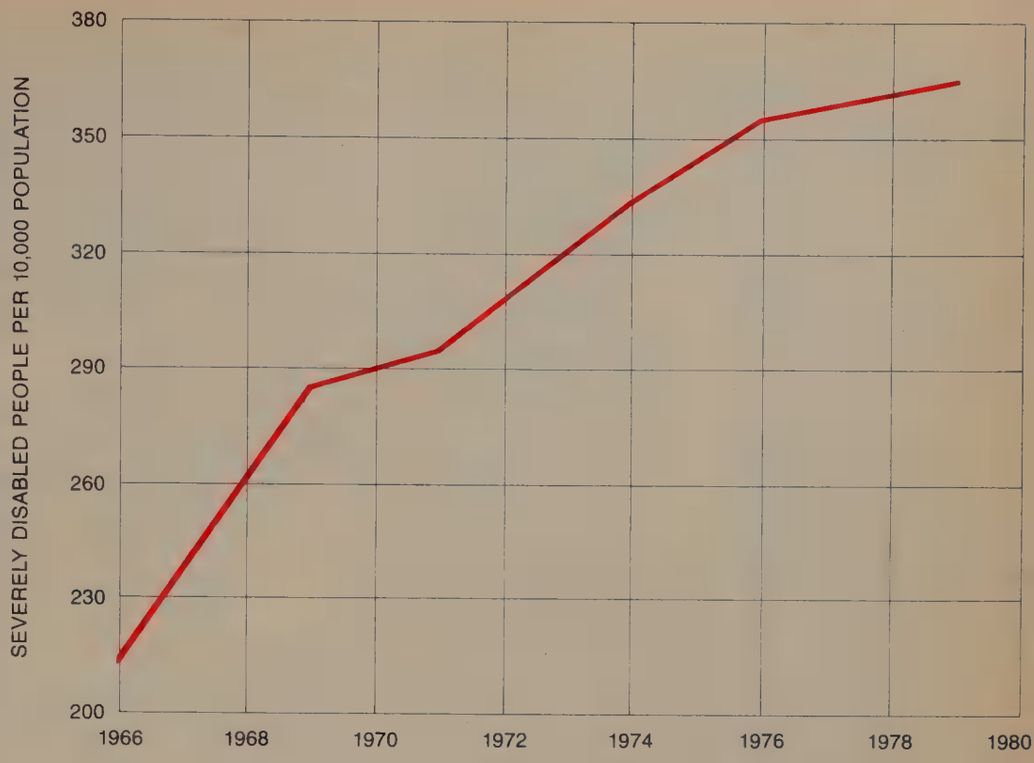
barrier-free movement, there is in one sense no turning back. The groundwork—accessibility standards, laws, regulations, technical know-how, public consciousness and, to a degree, professional awareness—on which the movement is based is substantial. Many of the gains are here to stay. In short, the efforts of the past two decades to develop accessibility standards and enforcement mechanisms have paid off. The built environment, in spite of many shortcomings, is far more accessible to mobility-impaired persons than it was a few years ago.

In considering the future of environmental accessibility for physically disabled Americans one must consider the full range of policy tools that can be used to stimulate public and private efforts toward achieving a barrier-free environment. The central weakness in the current policy of environmental accessibility is its overreliance on just one approach to policy implementation: the regulatory approach. When this approach is not accompanied by other incentives and inducements, it becomes inherently adversarial: it eventually pits one group against another. The adversarial nature of the regulatory approach has also helped to polarize the national debate on environmental accessibility and to reduce the chances for mutual accommodation.

Any critique of the regulatory approach, however, must take into account the fact that the development of accessibility regulations arose in direct response to the failure of voluntary compliance. To repeal existing regulations would be to return to an earlier era of neglect. The champions of voluntarism would do well to consider this earlier period in the history of the barrier-free movement. Although the development of the first accessibility standards had considerable private sponsorship, the standards themselves had little private endorsement in the form of active implementation.

Seldom can a public-policy goal such as environmental accessibility be obtained with the benefit of just one policy instrument. Other approaches must be tried. Two are noteworthy here: technical assistance and economic incentives.

In our review of the architectural-barriers problem we observed how mediating bodies providing technical assistance to designers, builders, clients and users can help to minimize unnecessary conflict. Inherent in the technical-assistance approach to policy implementation is a large measure of mutual problem solving and cooperation. The need for technical assistance is recognized in the statutory provisions governing the Architectural and Transportation Barriers Compliance Board. So far, however, funding for technical assistance at the Federal level has been quite modest.



INCIDENCE OF SEVERE DISABILITY has increased by more than 70 percent over the 13-year period from 1966 to 1979, rising from 213 to 365 Americans per 10,000 population.

One of the most powerful tools of policy implementation is the variety of economic incentives that government offers through the tax system and through government expenditures. The economic-incentives approach to policy implementation has not been widely tried with respect to architectural barriers. Some states have introduced modest tax deductions for the removal of such barriers. Until recently the Federal income-tax code allowed businesses a modest tax deduction of $25,000 in any given year. The statutory authority for the Federal tax deduction, however, expired last year. Although the size of the Federal tax write-off was far too small to be noticed by most large businesses, such cost-sharing schemes can help to foster compliance with governmentally sponsored accessibility standards.

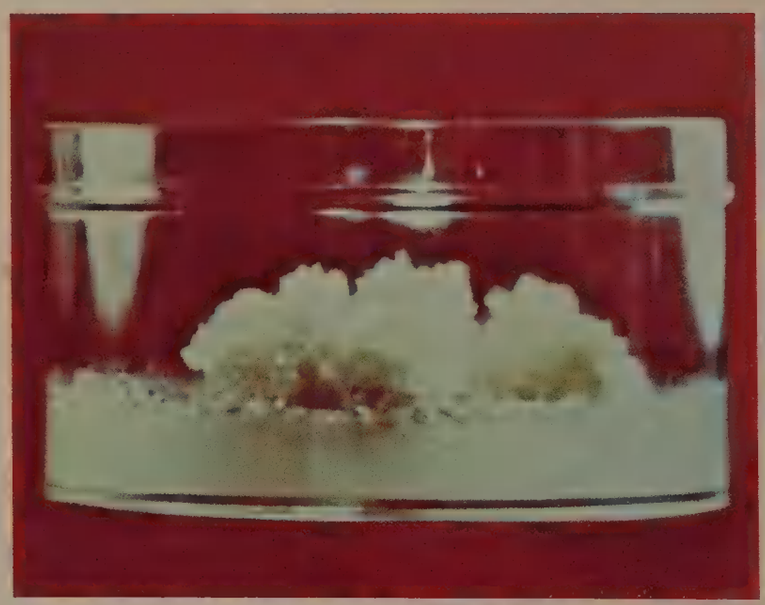
Pervading the public debate on environmental accessibility is a legal standard commonly referred to as the "reasonable accommodation" rule. This rule was affirmed in 1979 by the United States Supreme Court in the case of *Southeastern Community College* v. *Davis,* in which the court ruled that "substantial modifications" resulting in "undue administrative or financial burdens" were not required under Section 504 of the Rehabilitation Act of 1973.

Implicit in the reasonable-accommodation rule is a cost-benefit consideration: Are overall costs reasonable in the light of anticipated benefits? One should be cautious, however, about shaping disability policy solely on the basis of cost-benefit analysis. In the extreme there are instances where taking cost-benefit criteria to their logical conclusion could entail withholding essential life-support systems. We believe in many cases a more appropriate decision rule would be the cost-effectiveness criterion: How, in the face of limited resources, can a particular right be honored or societal responsibility be met in the least costly or most cost-effective way? Phrasing the issue in these terms does not challenge the rights of disabled persons or absolve society of its responsibilities, but it does face up to the unavoidable economic consequences.

The increasing preoccupation with cost-benefit analysis also betrays the degree to which society hesitates to make financial outlays on behalf of disabled people. The current outlook on disability issues would change materially if it were clearly understood, as we have argued, that much of the cost is the product of society's success over the years in saving and prolonging the lives of millions of Americans, often at the cost of a long-term residual disability. This success has its price, and it is a price that must be paid. For disabled people environmental accessibility is fundamental to their quality of life and their ability to participate in the life of the community.

The ultimate and most pervasive of environmental barriers are the attitudinal ones, particularly the view that disabled people are helpless, pathetic victims deserving of charitable intervention. There is now more than enough experience to indicate that disabled people can, with appropriate environmental supports, lead full and independent lives. Without the removal of attitudinal barriers the disability legislation of the past decade will not realize its full promise.

CROWN GALL on a tobacco plant has developed where the top of the plant was cut off and the cut stem was inoculated with *Agrobacterium tumefaciens*. The infecting bacteria carry a circular molecule of accessory DNA: the *Ti* (tumor-inducing) plasmid. A segment of the plasmid called *T*-DNA (transferred DNA) is inserted into plant cells. It gives rise to the tumor, which in this case is a teratoma: a tumor mass from which shoots and leaves grow in an unorganized way. Geneticists hope to exploit the plasmid as a vector for introducing genes into plants, giving rise to plants with improved traits. The photograph was made by Robert and B. Gillian Turgeon of Cornell University.



TOBACCO TUMOR TISSUE grows in culture in the author's laboratory at Washington University in St. Louis. The bacteria have been eliminated from these cell lines derived from infected plants, but their influence remains as a genetic change in the transformed plant cells. The cells grow in culture without any added auxin or cytokinin, hormones needed to enable normal plant cells to grow in culture. Here two kinds of tumorous growth are shown: a teratoma at the left and a callus (a mass of undifferentiated cells) at the right.

A Vector for Introducing New Genes into Plants

The induction of a plant tumor by a bacterium is a natural form of genetic engineering. The piece of DNA the bacterium injects may serve as a tool for the genetic modification of crop plants

by Mary-Dell Chilton

For a long time, perhaps millions of years, the common soil bacterium *Agrobacterium tumefaciens* has been doing what molecular biologists are now striving to do. It has been inserting foreign genes into plants and getting the plants to express those genes in the form of proteins. In the process the bacterium causes the plant cells to proliferate and form a gall, or tumor, most commonly at or near the junction of the root and the stem (the crown).

Crown gall disease can afflict a wide range of dicotyledonous (broad-leaved) plants, and it causes considerable loss in certain crops, notably grapes, stone fruits and ornamental plants. Since the turn of the century it has been studied by plant pathologists because of its economic impact. In the past few years, as it has become clear that the infective process is a natural form of genetic engineering, *A. tumefaciens* and the tumor it engenders have become the subject of fiercely competitive investigation by molecular geneticists in both academic and industrial laboratories. What has long appeared to be simply the agent of a bothersome plant disease is likely to become a major tool for the genetic manipulation of plants: for putting new genes into plants and thereby giving rise to new varieties with desired traits.

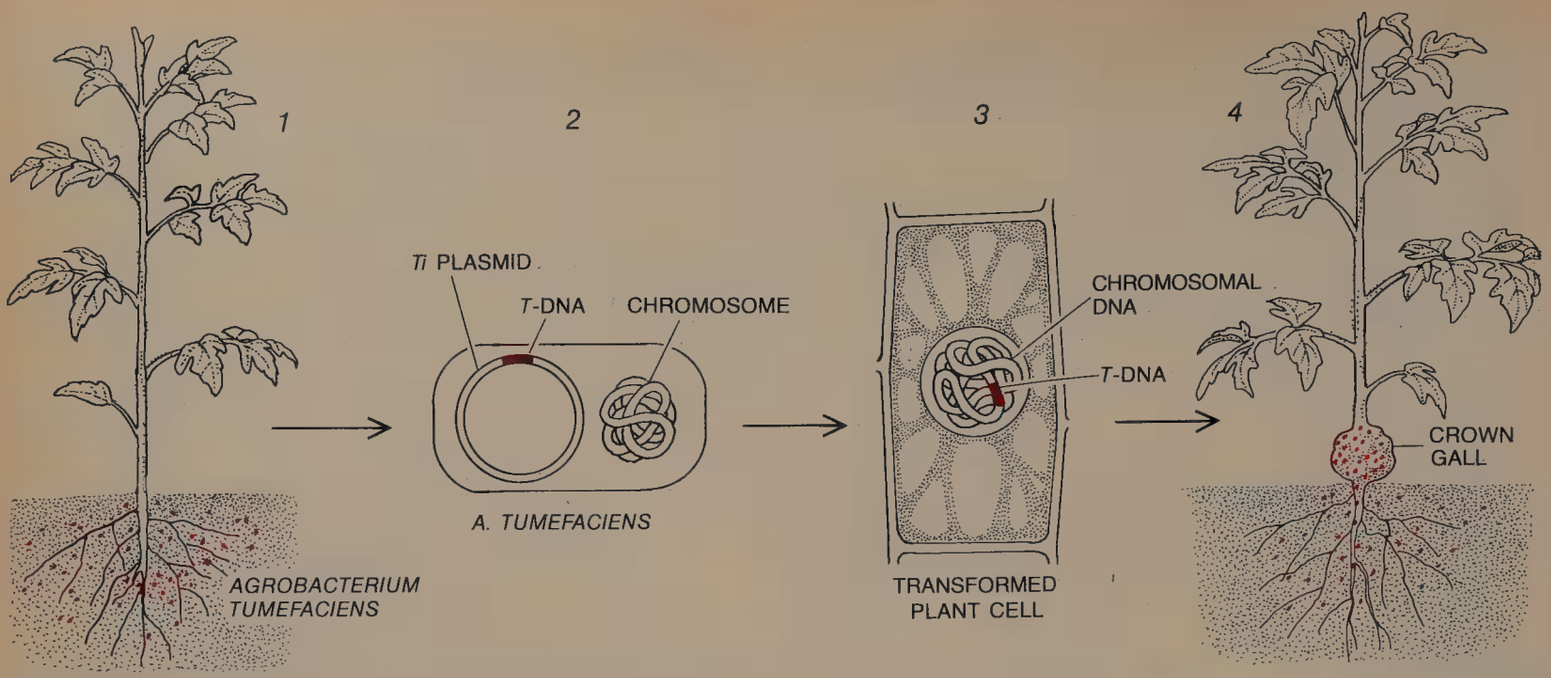
In 1907 Erwin F. Smith and C. O. Townsend of the U.S. Department of Agriculture identified *A. tumefaciens* as the causative agent of crown gall disease. They did so by isolating the bacteria from plant galls and then showing that inoculation of the bacteria into a wound in a healthy plant would generate a new gall. It has since become clear that infection takes place only in freshly wounded tissue. The wound site becomes refractory after only two days or so; indeed, washing the "wound juice" (the sap that flows out of cut tissue) away from the site of a fresh wound can block the induction of a tumor. Not all strains of *A. tumefaciens* are virulent, or capable of inducing a tumor. Virulent bacteria are known to bind to specific pectin constituents of the plant-cell wall. The bacterial cells actually enter only dead, broken plant cells, and yet they stimulate adjacent living plant cells to form a tumor; the bacteria themselves proliferate among the tumor cells.

In 1947 Armin C. Braun of the Rockefeller Institute for Medical Research first succeeded in cultivating gall tissue, free of the inducing bacterium, in the laboratory on a completely defined medium containing only sucrose and inorganic salts. He found that the tissue had the characteristic property of a tumor: a habit of uncontrolled growth. Whereas normal plant cells or tissues grow slowly in culture, and then only when plant hormones (cytokinins and auxins) are supplied, the gall tissue grew rapidly in the absence of any exogenous hormone. It continued to grow luxuriantly when it was transferred to fresh medium time after time. Braun concluded that the plant cells had somehow been transformed—changed to tumor cells—by a hit-and-run act on the part of *A. tumefaciens*. He proposed that the bacterium introduces a "tumor-inducing principle" into the plant cell.

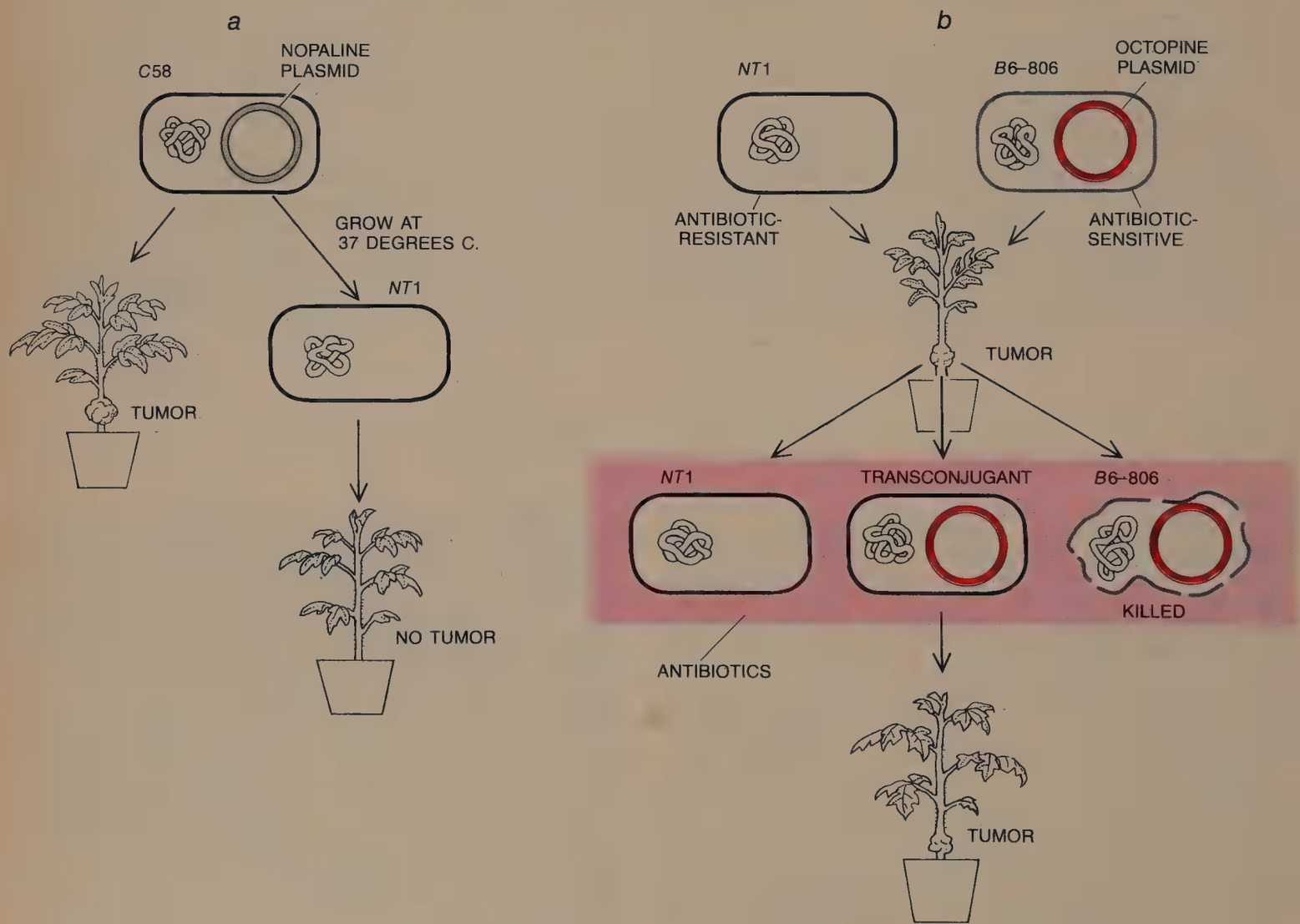
An important metabolic clue to the nature of the tumor-inducing principle was discovered in the 1960's by Georges M. Morel and his colleagues at the Institut National de la Recherche Agronomique in Versailles. They found that crown gall cells synthesize a class of novel chemical compounds (which they named opines) that are not present in normal cells of the same plant. The opines are derivatives of common metabolic intermediates, for the most part amino acids (the constituents of proteins) and various keto acids or sugars. Two opines were studied intensively: octopine, a compound of the amino acid arginine and pyruvic acid, and nopaline, a product of arginine and alpha-ketoglutaric acid. Morel's group established two significant findings. One was that whether a tumor synthesizes octopine or nopaline depends not on the species of the host plant but on the strain of the bacterium that induces the tumor. The other was that a given bacterial strain is able to grow on either octopine or nopaline but not on both; it can metabolize whichever of the compounds is synthesized by the tumor it induces. Morel made a striking prediction. He proposed that the bacterium must insert into the plant cell a gene governing the synthesis of either octopine or nopaline; the tumor-inducing principle must be DNA.

If that is the case, it soon became clear, the DNA must be on some mobile element. Allen Kerr of the University of Adelaide found that the property of virulence was transferred from a virulent strain to an avirulent one when the two strains were inoculated into the same plant. The transfer took place even between distantly related species of *Agrobacterium*, making it unlikely that the genes responsible for virulence were on the bacterial chromosome. Kerr thought the genes were probably carried by a bacterial virus or a plasmid: a small circle of nonchromosomal double-strand DNA, usually carrying ancillary genes such as those affecting pathogenicity, the metabolism of rare compounds or resistance to antibiotics.

Additional evidence that virulence in *Agrobacterium* might be determined by a virus or a plasmid came from the finding that virulence could be lost. In 1971 Robert H. Hamilton, Jr., and M. Z. Fall of Pennsylvania State University discovered a strain of *A. tumefaciens* whose virulence trait is temperature-sensitive. When strain *C*58 is grown at a moderately high temperature (37 degrees Celsius), the bacteria lose virulence; they never spontaneously regain the ability to induce a tumor. In other words, something is permanently lost from the cells. Again the implication was that the tu-



TUMOR IS INITIATED when bacteria enter a lesion, which is usually near the crown of the plant (the junction of root and stem), and attach themselves to cells (*1*). A virulent bacterium carries, in addition to its chromosomal DNA, a *Ti* plasmid (*2*). The plasmid's *T*-DNA is introduced into a cell and becomes integrated into the cell's chromosomal DNA (*3*). Transformed cells proliferate to form a crown gall tumor (*4*). The tumor cells synthesize compounds called opines, which serve as nutrients for *A. tumefaciens* cells inhabiting the gall. Two well-studied opines are octopine and nopaline. The *Ti* plasmid carried by a given strain of the bacterium induces synthesis of a given opine.



PLASMID IS IDENTIFIED as the tumorigenic agent by experiments such as these. *A. tumefaciens* strain *C*58 is normally virulent (*a*). It grows on nopaline and induces tumors that synthesize nopaline. After growth at 37 degrees Celsius, *C*58 loses virulence and the ability to grow on nopaline; the *Ti* plasmid of *C*58 is missing from the avirulent derivative, *NT*1. Virulence can be restored to *NT*1 by mating with a virulent strain (*b*). *NT*1 cells (mutants resistant to two antibiotics) are inoculated into a tomato plant with *B*6–806, a virulent strain that grows on octopine. A tumor forms. It is ground up and spread on a medium with the two antibiotics, so that only resistant *NT*1 cells form colonies. Many of the cells are transconjugants: they carry the octopine-type plasmid of strain *B*6–806, which they have acquired through conjugation, a bacterial version of mating. Such cells are virulent, grow on octopine and induce octopine-synthesizing tumors.

mor-inducing genes must be on a virus or a plasmid.

The elusive DNA was identified in 1974. Jeff Schell and Marc Van Montagu and their co-workers at the State University of Ghent found very large plasmids in all virulent strains; the plasmids could not be found in avirulent strains. Soon it was shown that the loss of virulence by strain *C*58 at 37 degrees C. is due to the loss of its plasmid. The plasmid is Kerr's mobile element; it is transferred from a virulent strain to an avirulent one by conjugation, a bacterial version of mating. Conjugation experiments done by Rob A. Schilperoort and his colleagues at the State University of Leiden revealed it is the plasmid that confers on bacteria the ability to catabolize (break down) octopine or nopaline, and so to grow on one of those compounds, and that determines whether the tumor will synthesize octopine or nopaline. Even the spread of a particular virulence plasmid through a population of *A. tumefaciens* is promoted by the presence of the opine whose synthesis and catabolism it specifies: the transfer of the plasmid in conjugation is specifically induced by a particular opine.

The *Ti* (for tumor-inducing) plasmid, in other words, is the central element of a rather wonderful, highly evolved ecological interrelation. The presence of a *Ti* plasmid enables a bacterium to subvert the plant cells' metabolic machinery to make a substance that is of no use to the plant but on which the bacteria (and hence also the symbiotic plasmid) can proliferate. The growth of a tumor makes for a large supply of that substance. In nature the tumor tissue is at soil level, and its opine presumably diffuses into the soil, where the transfer of plasmids to hitherto avirulent bacteria promotes the survival of both the bacterial species and the plasmid.

If a crown gall is generated by gene transfer and if the genes are on the *Ti* plasmid, clearly plasmid DNA should be present in crown galls. At the University of Washington in Seattle, Eugene W. Nester, Milton P. Gordon and I, together with a number of postdoctoral and graduate students, undertook to find that DNA. At the time the best way to detect a particular DNA (the *Ti* plasmid in our case) in the presence of a large amount of other DNA (the plant's own) was to measure the rate of DNA reassociation. The DNA double helix is denatured (its two complementary strands are separated) by heating to 100 degrees C. When the temperature is reduced to 68 degrees, complementary strands that "find" each other renature, or reassociate, to form a double helix again. The rate of reassociation is proportional to the square of the concentration of the DNA. Our approach was to compare the rate of reassociation of denatured *Ti* plasmid DNA alone with the rate of its reassociation in the presence of normal plant DNA or in the presence of crown gall DNA. If the tumor-cell DNA included copies of the plasmid DNA, it should add to the concentration of complementary strands, and so make the plasmid DNA (which we could detect because it was labeled with a radioactive isotope) renature faster, in a quantitatively predictable way.

Our initial results showed that the addition of denatured tumor DNA to denatured plasmid DNA made no significant difference in the reassociation rate, so that clearly there was no copy of the entire *Ti* plasmid in the tumor. Might part of it be there? We cut up the plasmid DNA with a restriction endonuclease: one of a family of enzymes, each of which recognizes a particular short sequence of nucleotides (DNA subunits) as a unique cleavage site and therefore always cuts any given DNA into a reproducible set of fragments. The endonuclease *Sma* I cuts a *Ti* plasmid into some 20 fragments of different sizes, so that digesting a quantity of plasmid DNA yields some 20 unique kinds of fragments. These can be separated according to their size by gel electrophoresis. Each group of fragments forms a narrow horizontal band on the gel; the bands can be visualized by staining them with a fluorescent dye and the DNA in each band can be purified for further experiments.

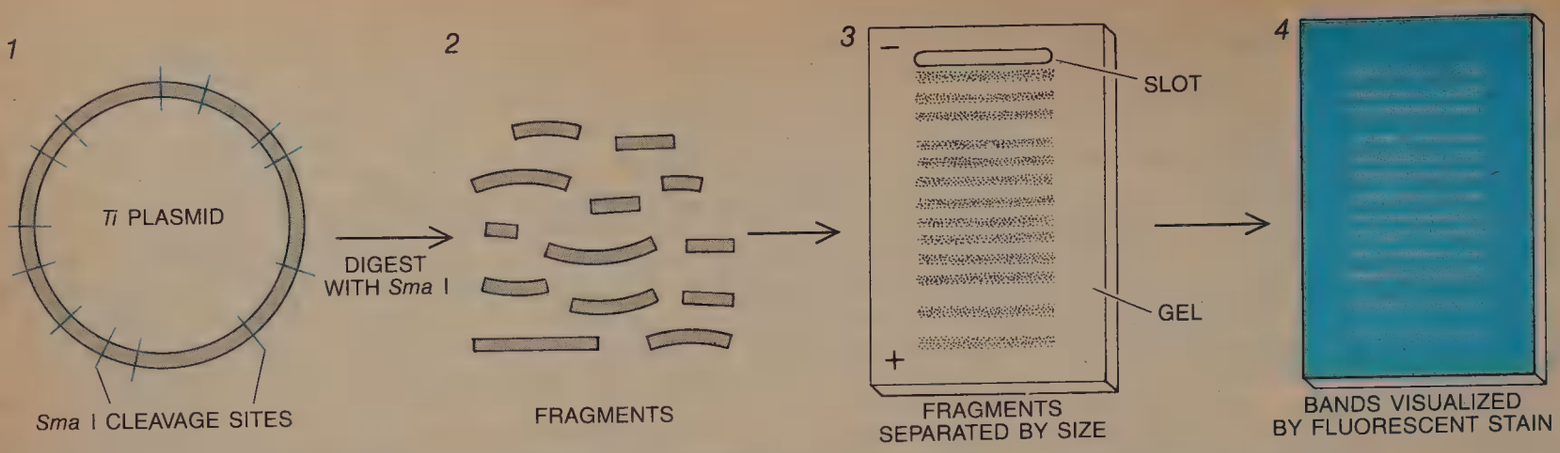
When we subjected the *Ti* plasmid DNA to electrophoresis and then did the renaturation experiment with the individual DNA fragments instead of the plasmid as a whole, we were excited to find that two of the fragments renatured faster in the presence of tumor DNA; the others did not. Those two segments of the *Ti* plasmid must therefore be present in the tumor cells. Subsequent experiments with many tumor lines have shown that specific fragments of the *Ti* plasmid that induced the tumor are always present in the crown gall cells. We call these fragments transferred DNA, or *T*-DNA.

Analysis of renaturation rates has now been supplanted by a more direct method of detecting particular segments of DNA embedded in a large amount of nonspecific DNA. Southern-blot analysis, developed by E. M. Southern of the University of Edinburgh in 1975, combines electrophoresis with renaturation, and it has become a fundamental tool of molecular genetics. To find plasmid DNA in tumor cells one begins by digesting the tumor-cell DNA with, for example, *Sma* I. Whereas such a digest yielded 20 fragments from each plasmid, however, it produces a million or so fragments of the enormously larger genome (the total gene complement) of the plant-tumor cell. When these are spread out on a gel by electrophoresis, they form what looks like a continuous smear (although it is really a set of a million or so discrete bands). Any *T*-DNA fragments in the tumor are buried in that smear; the Southern technique picks them out.

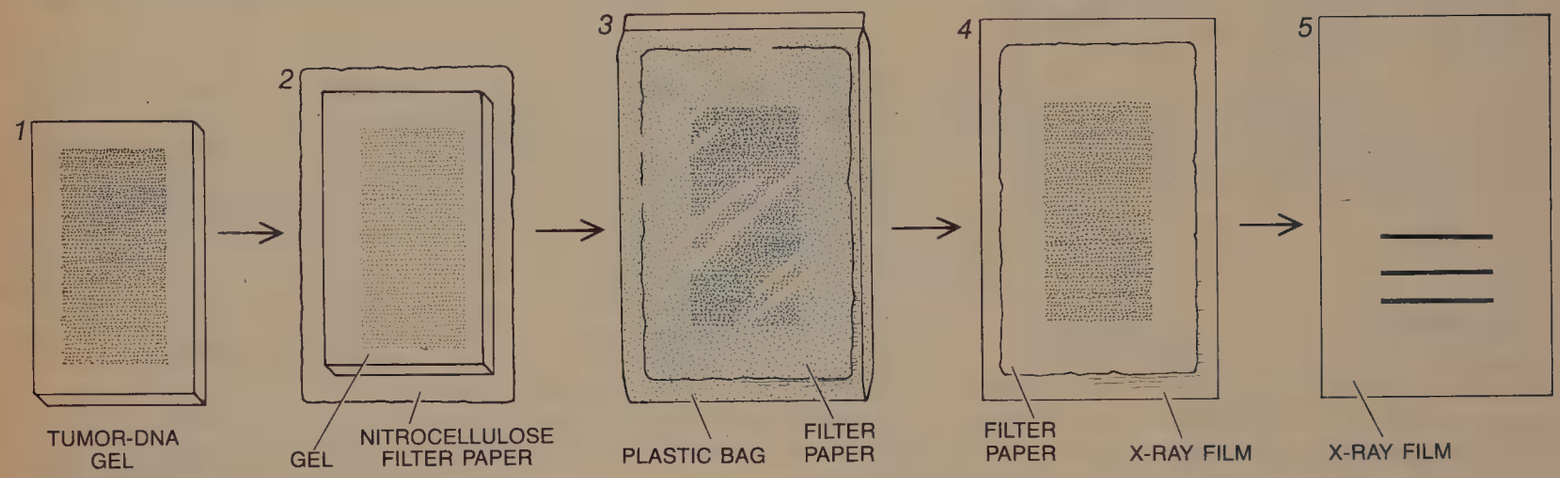
The tumor DNA fragments are denatured and then transferred by blotting to a sheet of nitrocellulose filter paper, which now carries a sharp but invisible print of the original pattern. The DNA is fixed to the paper. The paper is incubated with a solution of denatured, radioactively labeled fragments of plasmid DNA, which serve as a probe. If there is plasmid DNA in the tumor, probe DNA reassociates with it, forming hybrid double-strand DNA; the rest of the probe DNA is washed away, leaving radioactive plasmid DNA bound to its counterpart strands in the tumor DNA. The paper is put in contact with X-ray film. When the film is developed, dark bands on the autoradiogram show where *T*-DNA has been incorporated into plant DNA. Comparison of those bands with the bands on a gel displaying a *Sma* I digest of pure plasmid DNA shows which *Sma* I fragments are, in whole or in part, in *T*-DNA: fragments 3*b*, 10*c*, 16 and 17.

To understand the structure of *Ti* plasmids and their *T*-DNA in detail it was necessary first to determine how the plasmid's fragments map into a closed circle and then to locate genes for particular functions on the circular fragment map. Our group at the University of Washington deduced the first fragment map of a *Ti* plasmid (for the octopine plasmid designated *B*6–806) by digesting the plasmid with the endonuclease *Hpa* I and seeing how the resulting fragments overlapped with the *Sma* I fragments cut from the same plasmid. We prepared a large Southern blot of the *Hpa* I bands [*see illustration on page 55*] and cut it into longitudinal strips. A different labeled *Sma* I fragment was allowed to hybridize with each strip. A labeled fragment that binds to only one *Hpa* I band must map within the band; a fragment that binds to two *Hpa* I bands must span the junction between those bands, and so on. By applying criteria such as these and by testing fragments generated with several endonucleases one finally builds up a complete fragment map.

The next step was to map the plasmid's genes by associating mutations at particular sites with changes in particular functions. For this purpose several groups exploited the technique of transposon mutagenesis. Some genes for resistance to an antibiotic are carried on transposons, which are short segments of DNA that tend to jump, almost at random, to new sites on the bacterial chromosome or on a plasmid. The presence of a transposon simultaneously



Ti PLASMID IS ANALYZED by gel electrophoresis. The plasmid is digested with a restriction endonuclease, an enzyme that cleaves DNA at specific sites. Fragments made by the endonuclease *Sma* I (actually thousands of copies of each fragment, from thousands of copies of the plasmid) are placed in a slot at one end of an agarose gel. In an electric field the negatively charged DNA fragments move toward the positive pole. The smaller they are, the faster they move, and so the fragments form a series of bands, which can be visualized with a fluorescent dye that binds to DNA. A given plasmid digested with a given endonuclease yields a distinctive pattern of bands.



PRESENCE OF PLASMID DNA (*T*-DNA) in gall tissue is demonstrated by Southern transfer. Tumor-DNA fragments generated by digestion with *Sma* I are separated by gel electrophoresis. The thousands of bands of fragments form a continuous smear on the gel (*1*). The DNA is denatured (the two strands are separated), transferred by blotting to filter paper and fixed to it (*2*). The paper is incubated with probe DNA (*color*): *Ti*-plasmid DNA or fragments of it, labeled with a radioactive isotope and denatured (*3*). Probe fragments reassociate with complementary DNA, that is, they "find" and bind to bands constituting *T*-DNA. Unbound probe DNA is washed away (*4*). The *T*-DNA bands are made visible by autoradiography (*5*). The fragments can be identified by comparison with the original *Sma* I gel.

generates a mutation (because insertion of the transposon interrupts a gene on the chromosome or plasmid), identifies itself (because the bacterium is rendered resistant to an antibiotic) and flags its location (because the pattern of endonuclease fragments is changed by the insertion of new DNA into an old fragment, giving rise to new ones). At the University of Washington, Nester and David Garfinkel introduced the transposon by means of a "suicide vector": a plasmid, constructed by John Beringer of the John Innes Institute in England, that carries the kanamycin-resistance transposon *Tn*5. Introduced into *Agrobacterium,* this plasmid conveniently self-destructs, but only after the *Tn*5 transposon has had an opportunity to jump to a new site on the bacterial chromosome or on a *Ti* plasmid.

One selects kanamycin-resistant bacterial colonies in which the *Tn*5 is on the plasmid rather than on the chromosomes [*see top illustration on page 56*] and then tests them for any change in genetic properties. Many mutations are "silent," but in some cases there is a detectable change. If the bacteria cannot catabolize octopine, for example, a gene needed for that process must have undergone mutation. To map the gene one isolates *Ti* plasmid DNA from the mutant bacteria, digests it with *Sma* I and examines the resulting electrophoresis pattern. It is different from the normal pattern: one of the usual fragments (call it *X*) is missing and two new fragments are seen, and the combined size of the two new fragments equals the size of fragment *X* plus *Tn*5. The gene must be within fragment *X*. Knowing that *Sma* I cuts near the middle of *Tn*5, one can determine, from the size of the two new fragments and certain other information, the precise location of the gene.

From this kind of analysis by Nester's group, by Schell and Van Montagu and by Schilperoort's group, a genetic map of octopine and nopaline *Ti* plasmids eventually emerged. A large sector of the plasmid to the left of the *T*-DNA carries the virulence functions. Mutations in this sector presumably block the attachment of the bacterium to a plant cell, the transfer of plasmid DNA into the cell or the integration of the DNA in the plant-cell chromosome, or perhaps several of these functions. Another large sector, on the other side of the *T*-DNA, governs the uptake and catabolism of octopine or nopaline. Genes governing the plasmid's ability to transfer in conjugation have been mapped, as have genes for the catabolism of additional opines. Large sectors of the plasmid still appear to be silent, but that may be because they govern traits of which we are not yet aware, perhaps including the ability to catabolize still more nutrient substances.

Mutations within the *T*-DNA were of particular interest in that they identified genes transferred from the bacterium to the plant. In the nopaline plasmid studied by Schell and Van Montagu one mutation eliminated nopaline synthesis in tumors, and it was situated within the *T*-DNA. Morel's prescient hypothesis was thereby confirmed; here was direct evidence, at the level of the gene, that the ability to synthesize nopaline is governed by DNA transferred from the

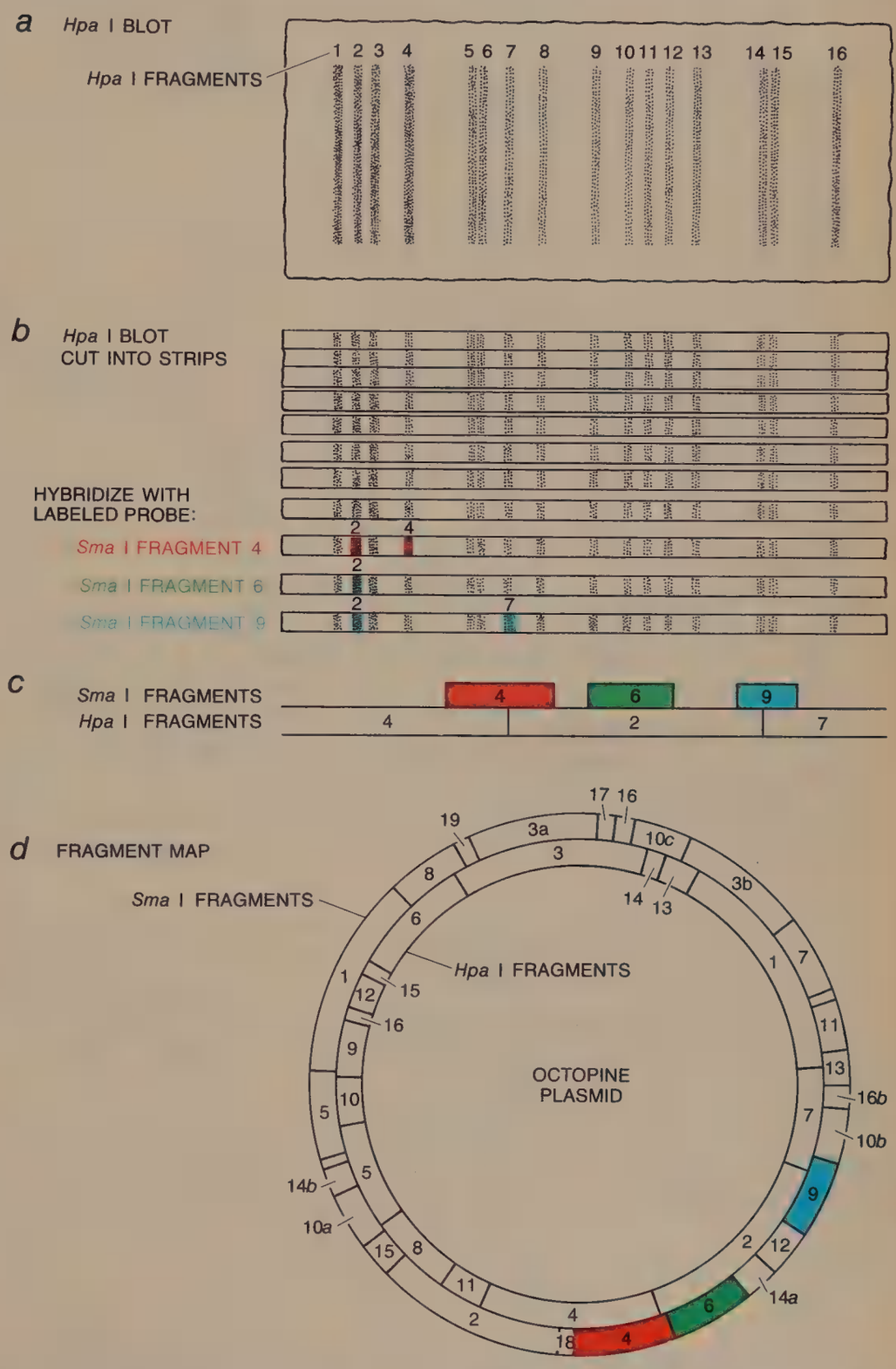
bacterium to the plant. Cells of the tumor incited by this mutant *Ti* plasmid turned out to harbor not only the usual *T*-DNA but also a copy of the transposon that had given rise to the mutation. Now, the transposon in question, *Tn*7, is a sizable piece of DNA. Its presence in the tumor cells meant that a large segment of foreign DNA, enough to carry several genes, can be transferred along with the *T*-DNA. Clearly, then, in principle it should be possible to make *T*-DNA a vector for introducing a desirable gene into plant cells. If such a gene could be made to function in the plant cells and if the cells could be regenerated into healthy plants, a method would be available for genetic engineering in plants.

Additional transposon mutations in *T*-DNA enabled Schilperoort's group to identify sites controlling the perturbation of hormone activity that is apparently responsible for tumor growth. The mutations bring about two kinds of changes in tumor morphology. One class of mutations cause roots to sprout from the tumor and the other class give rise to shoots. Just such changes are caused in normal plant tissue by a hormone imbalance: a high ratio of auxins to cytokinins induces excess formation of roots and a low ratio of auxins to cytokinins generates more shoots. Schilperoort showed that the effect of the tumor-morphology mutations could indeed be reversed by correcting the putative hormone deficiency, that is, by adding cytokinins to the medium in the case of "rooty" mutants and by adding auxins in the case of "shooty" mutants.

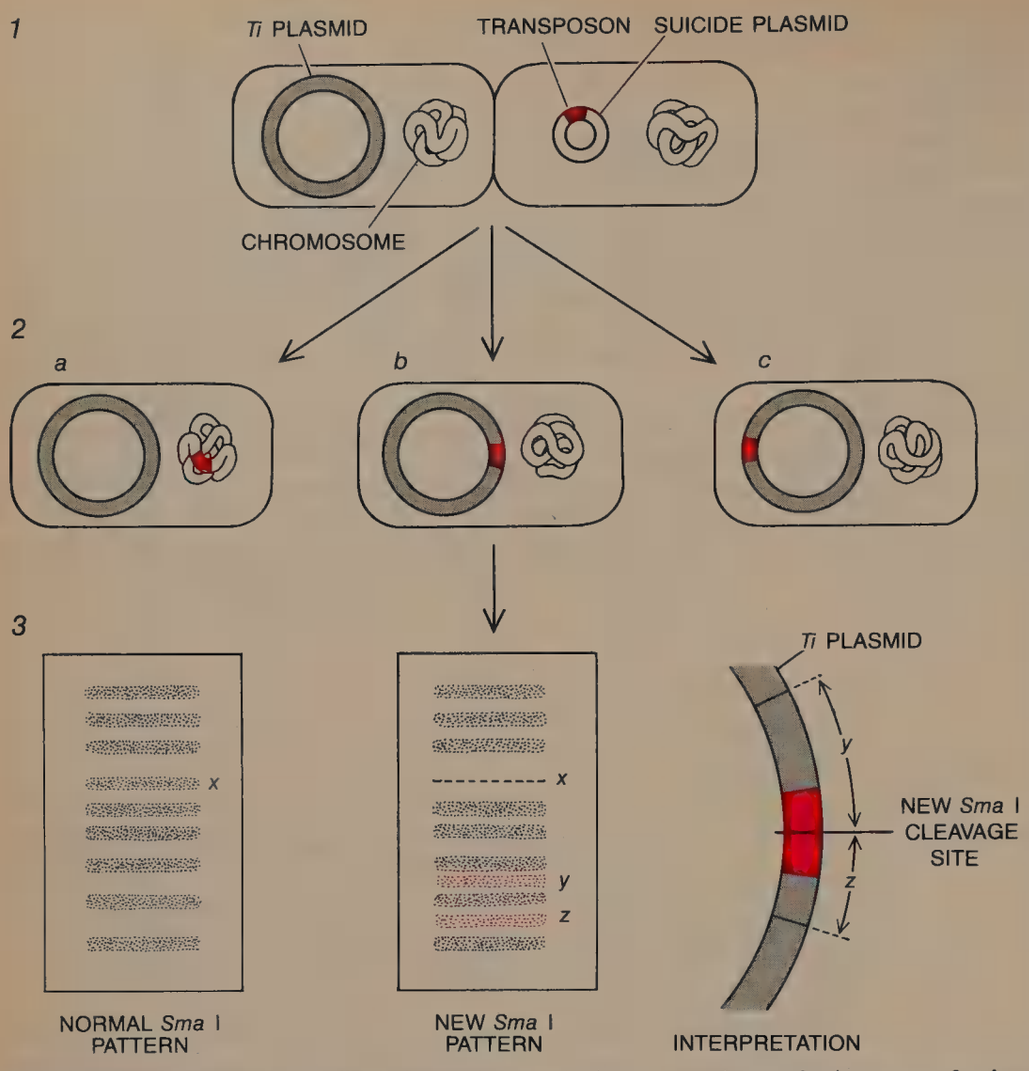
In Nester's laboratory several students worked out a detailed map of *T*-DNA, including the rooty and shooty gene loci. Again the method was transposon mutagenesis, but now that something was known about the structure of the *Ti* plasmid the mutagenesis could be "site-directed": one could choose a particular *T*-DNA fragment, clone it in a small vector plasmid, generate a mutation in it and then transfer the fragment into *A. tumefaciens* and onto the *Ti* plasmid by means of double recombination. The result was a *Ti* plasmid with a mutation in a predetermined segment.

Intensive mutagenesis of *T*-DNA in this way defined the rooty and shooty loci and also the site of another class of mutations generating unusually large tumors. Extensive regions of the *T*-DNA appeared to be silent, that is, mutations within them had no apparent effect. Curiously, none of the 75 mutant plasmids studied had completely lost its virulence. This implies that all the genes governing the transmission of *T*-DNA to a plant cell and its insertion into the chromosomal DNA must fall outside the *T*-DNA, in the virulence region to the left of it. In this respect *T*-DNA is different from both viruses and transposons, both of which carry with them the vital functions allowing them to transfer to a new host.

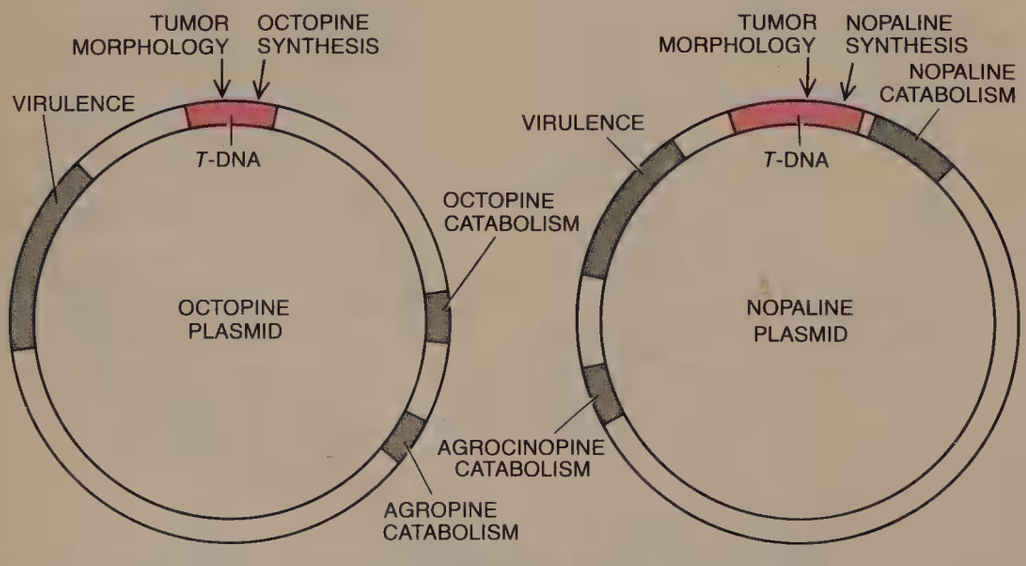
A gene is expressed when its DNA is transcribed into a messenger RNA and the RNA is translated into protein. Crown gall cells contain RNA transcribed from *T*-DNA genes. The RNA seems to have the special characteristics of messenger RNA in eukaryotic cells (the nucleated cells of all organisms higher than the bacteria) rather than those of prokaryotic (bacterial) messenger RNA. The sizes and map positions of individual molecules of the tumor RNA have been determined. Individual RNA's correspond to the rooty and the "large" tumor-morphology loci in the octopine and nopaline plasmids; two RNA's are encoded by the shooty locus. Interestingly, just about all the silent regions of *T*-DNA also encode messenger RNA's, some of which may correspond



FRAGMENT MAP OF PLASMID is derived by overlapping two sets of endonuclease fragments. A large Southern blot is prepared displaying the fragment pattern generated by cleaving the *Ti* plasmid with the endonuclease *Hpa* I (*a*). The blot is cut into strips and each strip is incubated with a different radioactively labeled *Sma* I probe, which reassociates with its *Hpa* I counterpart (*b*). *Sma* I fragment 4, for example, hybridizes with *Hpa* I fragments 2 and 4; *Sma* I fragment 6 hybridizes only with *Hpa* I fragment 2. Such data yield a linear pattern of overlapping fragments (*c*) and eventually yield a complete map of the circular *Ti* plasmid (*d*).



GENETIC MAP OF PLASMID is determined by transposon mutagenesis. A gene conferring antibiotic resistance may be carried on a transposon, a piece of DNA that tends to jump to a new site. A "suicide" plasmid carrying a kanamycin-resistance transposon (*color*) is introduced into *Agrobacterium* by conjugation (*1*). The transposon may jump (*2*) to a site on the bacterial chromosome (*a*) or on the *Ti* plasmid (*b, c*). Wherever the transposon becomes integrated it causes a mutation. By screening for kanamycin resistance it is possible to select cells in which the *Ti* plasmid has thus been subjected to mutation. Each cell is tested for a change in genetic function. The gene for that function can be determined by comparing the fragment pattern of the mutant plasmid with the normal pattern (*3*). Fragment *x* has disappeared from the plasmid of cell *b* and new fragments *y* and *z* are seen. They are the result of *Sma* I cleavage at a new site, in the transposon. The gene must be in missing fragment *x*, whose identity is confirmed by the fact that its size is equal to the sum of *y* and *z* minus the size of the transposon.



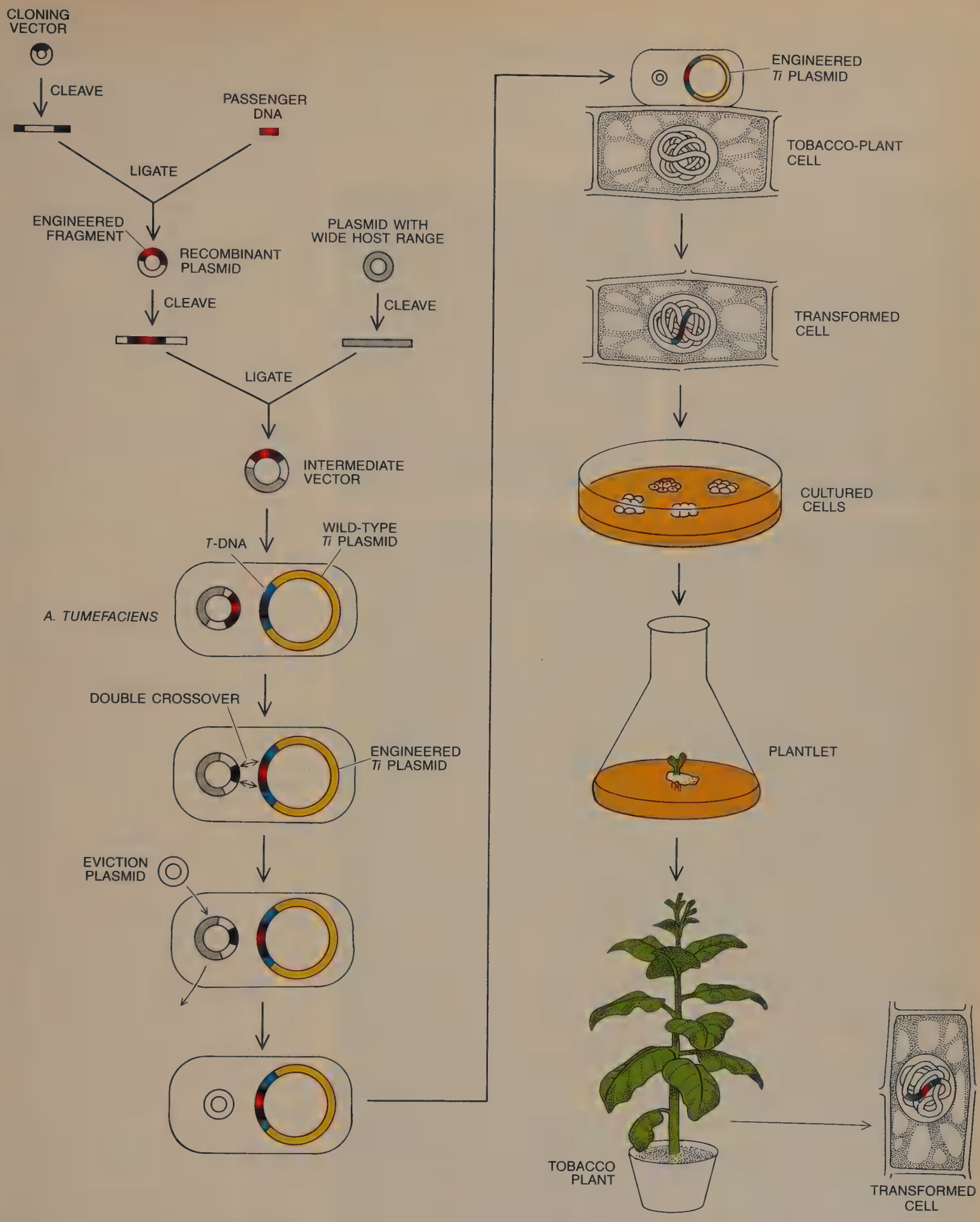
GENETIC MAPS are beginning to be built up for an octopine *Ti* plasmid (*left*) and a nopaline one (*right*). The *T*-DNA has been defined. Within it gene loci governing synthesis of octopine or nopaline have been mapped. So have genes affecting tumor morphology, which are revealed by the "rooty," "shooty" and "large" mutations. In both plasmids a region to the left of the *T*-DNA governs the virulence functions that enable the plasmid to induce a tumor. Regions outside the *T*-DNA also govern the catabolism of octopine or nopaline and other opines.

to genes for opines that have not yet been discovered.

The unique biological characteristics of the *Ti* plasmid make it a natural agent for gene transfer. It brings about the insertion into the plant's chromosomal DNA of *T*-DNA, which is thereupon maintained by the host plant, replicated along with the host plant's own genes, transcribed into RNA and (at least in large part) expressed as protein. How does the plasmid "decide" what part of itself is *T*-DNA? The answer may lie in specific signal sequences that have been identified in the border regions at both ends of the *T*-DNA. In these regions there is a 25-nucleotide "direct repeat" sequence of nucleotides. (Nucleotides are the four subunits of DNA whose sequence in the coding regions of a gene encodes the genetic information specifying a particular protein. The precise repetition of a short nucleotide sequence, either direct or "inverted," often serves as a signal for recombinational events.) The direct repeats may define the limits of the *T*-DNA; any foreign DNA inserted between the repeats should be handled as *T*-DNA, that is, it should be transferred to the plant cell.

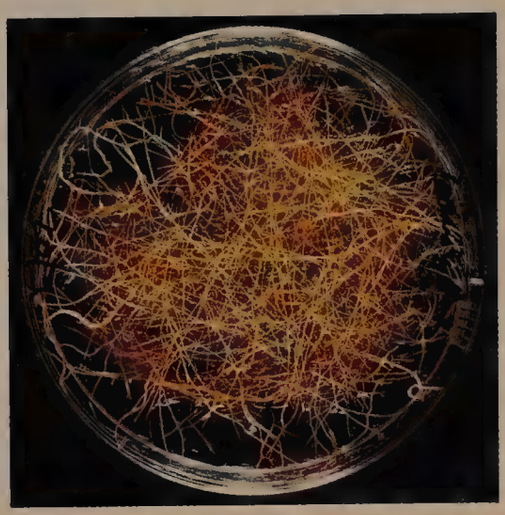
Let me describe one strategy for introducing a foreign gene into a plant by means of a *Ti* plasmid and *A. tumefaciens*. One begins with a recombinant plasmid: a small plasmid suitable for cloning in the common bacterium *Escherichia coli*, into which a known fragment of *T*-DNA has been spliced as in the site-directed mutagenesis process described above. The recombinant plasmid is cleaved open, at a site in the *T*-DNA segment, with a restriction endonuclease. A piece of "passenger" DNA is spliced into the opening. The passenger segment is bipartite: it consists of the foreign gene one wants to send into the plant on *T*-DNA and, attached to that gene, a genetic marker, say a bacterial gene for resistance to kanamycin. The combined plasmid is then recloned into a larger plasmid, which is introduced into an *Agrobacterium* strain carrying a wild-type (unmodified) *Ti* plasmid. As the bacteria proliferate a rare double-recombination event will sometimes take place, yielding a clone of bacteria whose *T*-DNA harbors an insert: the engineered passenger DNA. Such bacteria can be identified and selected by their survival on a medium containing kanamycin after the recombinant vector plasmid has been evicted. The selected bacteria are introduced into a plant. Their *T*-DNA is inserted into plant cells, and with it the passenger DNA.

A number of experiments have made it clear that this strategy or any of several similar ones can give rise to genetically modified crown gall cells growing in a culture dish. That result is still a far cry from the objective of growing genetical-



FOREIGN GENE IS INSERTED into a plant cell and a plant is regenerated. A small plasmid carrying a fragment of *T*-DNA (*black*) including the rooty locus is cut open with an endonuclease. Into it is inserted a piece of "passenger" DNA: a yeast gene for the enzyme alcohol dehydrogenase together with a gene for resistance to kanamycin (for screening). The recombinant plasmid is recloned in a plasmid that has a wide host range; the new intermediate vector is introduced into an *A. tumefaciens* cell carrying a wild-type (unmodified) *Ti* plasmid. The engineered *T*-DNA fragment in the intermediate vector can recombine with the same fragment (*black*) in the *Ti* plasmid. A rare double-crossover recombination brings the engineered fragment, with its passenger, into the *Ti* plasmid. The vector is evicted by introducing an "incompatible" plasmid. Bacteria carrying the engineered fragment in their *T*-DNA are identified by resistance to kanamycin. Such bacteria are used to transform tobacco-plant cells. The *T*-DNA, with its passenger, is integrated into the plant's DNA, giving rise to rooty mutant cells. The cells yield healthy plants whose transformed cells contain *T*-DNA carrying alcohol dehydrogenase gene.

ly modified healthy plants from such a transformed cell line. Few complete plants have been grown from wild-type crown gall cells, and in every case they have turned out to be mutants most or all of whose *T*-DNA has been deleted. The conclusion is that the presence of *T*-DNA genes is incompatible with healthy plant growth, presumably because they perturb auxin and cytokinin levels. How then is one to regenerate a healthy plant with the *T*-DNA and its passenger integrated into the plant's chromosomal DNA?



CARROT CELLS transformed by *Ri* plasmids, which are carried by *Agrobacterium rhizogenes*, regenerate into healthy carrot plants. The plasmid's *T*-DNA transforms cells of a slice of carrot into rapidly growing hairy roots (*top left*). A single root placed in a culture medium proliferates (*top right*). When a root is grown on a solid culture medium to form a callus and callus is transferred to a liquid medium, it forms an embryo that then develops into a plant (*bottom*). Other kinds of plant can also be regenerated from their hairy roots. The *Ri* plasmid seems to be a naturally disarmed gene vector that can give rise to genetically modified healthy plants. These photographs were made by Jacques Tempé of the University of Paris at Orsay.

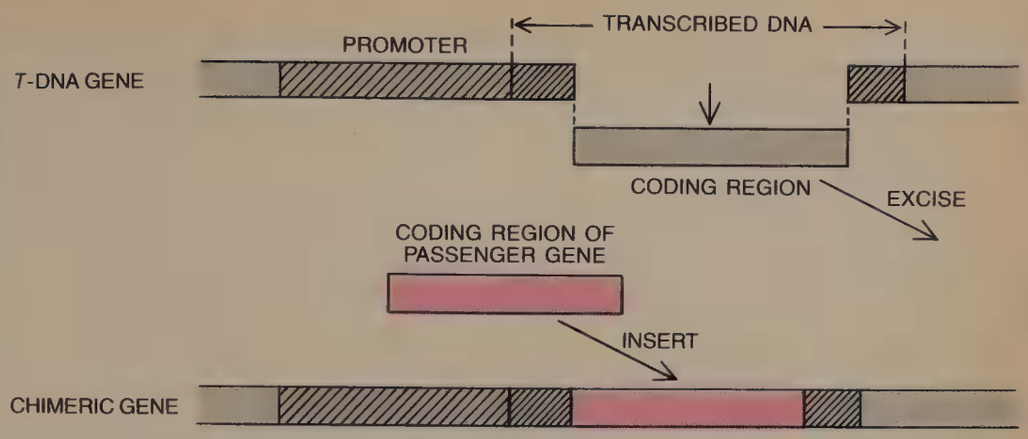
One possible way was suggested by some surprising results obtained recently by Andrew N. Binns of the University of Pennsylvania and by Kenneth A. Barton, Antonius J. M. Matzke and me in my laboratory at Washington University in St. Louis. We inserted a yeast gene encoding the enzyme alcohol dehydrogenase into the rooty locus of the nopaline plasmid and infected a tobacco plant with the resulting plasmid. The outcome was a culture of rooty mutant cells. The surprise was that the cells were not typically tumorous: they required added cytokinin for growth. Moreover, the cells (which synthesized nopaline and did not require added auxin) grew into complete tobacco plants. DNA isolated from the plants contained some 20 copies of *T*-DNA per cell—each including a copy of the inserted eukaryotic gene for alcohol dehydrogenase! In other words, far from being deletion mutants our plants had an all-time record number of copies of *T*-DNA per cell. After self-pollination the plants produced seeds that grew into healthy plants whose cells contained full-length *T*-DNA, still including the yeast gene, in multiple copies. We conclude that a rooty mutation somehow "disarms" nopaline *T*-DNA so that tobacco cells transformed by the *T*-DNA can regenerate. Whether this will work for other kinds of *Ti* plasmids and in other kinds of plants remains to be seen.

A more general way to build a disarmed *T*-DNA vector may be to delete all the *T*-DNA's own genes, leaving only the passenger DNA and the noncoding direct-repeat signal sequences. Such vectors have now been constructed in several laboratories, including mine. They may turn out to be the vectors of choice for most plants that are susceptible to infection by *A. tumefaciens*.

There is another family of *Agrobacterium* plasmids that may also serve as gene vectors yielding healthy genetically modified plants. They cause a proliferation of roots, called hairy-root disease, in plants infected by *A. rhizogenes* and are called *Ri* (for root-inducing) plasmids. The roots, like crown gall tissue, grow rapidly in culture in the absence of the infecting bacterium. Jacques Tempé and his colleagues at the Institut National de la Recherche Agronomique have shown that the root cells contain opines; they also contain several copies of a *T*-DNA. The important thing is that, like the rooty-mutant nopaline *Ti* plasmids, the *Ri* plasmids give rise to transformed plant cells that can regenerate into intact healthy, fertile plants. The *Ri* plasmids appear to be naturally disarmed vectors.

Given that foreign genes can be introduced into certain plants and that healthy genetically modified plants can be regenerated, the next question is

whether those genes will be expressed—will give rise to messenger RNA and protein—and so have an effect in their new setting. A number of experiments in which plant, animal, yeast and bacterial genes have been transferred have not been successful in this respect. Our yeast gene for alcohol dehydrogenase, for example, is not expressed (the enzyme is not synthesized) in the regenerated tobacco plants. The expression of a gene is regulated by various signal sequences flanking the DNA that codes for a protein. The most important regulatory sequences are in the promoter region, upstream from the gene. The signals surrounding an inserted foreign gene must be recognizable by the plant cell if the gene is to be expressed. For several years it has been possible to get bacteria to synthesize animal proteins by inserting into the bacteria a chimeric gene: one in which regulatory information from a bacterial gene is carefully spliced to the coding region of an animal gene.



CHIMERIC GENE must be constructed if a passenger gene is to be expressed by plant cells. The coding region of a *T*-DNA gene is excised precisely, leaving undisturbed the promoter region that precedes the gene (and regulates its transcription into RNA) and the signal sequences (*heavy hatching*) that flank the coding region (and may regulate the processing of the RNA for translation into protein). The coding region of the passenger gene is inserted between the regulatory regions so that it will be expressed as if it were an integral part of the *T*-DNA.

Very recently Michael W. Bevan in my laboratory, the group at Ghent and workers at the Monsanto Company in St. Louis have been able to make an analogous chimeric gene that is expressed in plant cells. In all three laboratories the procedure was to excise the coding region from a *T*-DNA gene, splice a bacterial gene for resistance to kanamycin between the *T*-DNA gene's regulatory regions and then introduce the chimeric gene into *Ti* plasmids more or less as described above. Plant cells transformed by the plasmids expressed the bacterial gene: they were resistant to a concentration of kanamycin that would have killed unprotected cells in culture. This success indicates that the problem of gene expression in a genetically modified plant can, in principle at least, be solved. As more is learned about the regulatory mechanisms of plant genes it should become possible not only to achieve expression but also to get selective expression by modifying the promoter region of genes. One might cause a gene to be expressed only in leaves or only in seeds, or only in response to a chemical signal.

So far both regeneration and expression have not been reported in the same experiment, but this will surely be accomplished soon. Then the question will be how to choose and isolate genes that are likely to improve crop plants. There are obvious objectives such as larger yields, resistance to pests and tolerance of cold and drought. The trouble is that there is no known way to find genes controlling such traits, most of which are probably governed by a number of genes and perhaps by elaborate regulatory mechanisms. Even if the appropriate genes could be identified, there is no known technique for isolating them by recombinant-DNA methods; only genes whose protein product is known have so far been isolated.

On the other hand, a feasible objective for the near future is the introduction into crop plants of novel foreign genes conferring resistance to herbicides or pesticides. If a plant can be grown in the presence of powerful agricultural chemicals, it should give a higher yield because of diminished competition from weeds or diminished predation.

The *Ti* and *Ri* plasmids' promise is limited to the modification of dicotyledonous plants, which *Agrobacterium* infects. The bacterium does not infect monocotyledonous plants, including grasses and cereals, and most of the world's important food crops are monocots. The reason for their immunity to the bacterium is not understood. There is as yet little information on how *T*-DNA travels from the plasmid to the plant cell's nucleus and on how (or even whether) *T*-DNA picks its sites of insertion into plant DNA. Research is needed in this important area, because better understanding of the sequence of events in infection may give insight into what step fails in the monocots. Many dicotyledonous plant species readily regenerate intact plants from tissue-culture cells; among the monocots regeneration is rarer and more difficult to achieve, but recently advances have been made in regenerating important crop plants from cells. That should spur efforts to extend the host range of the *Ti* plasmid to monocots and to develop new vectors that will function in monocots.

Agrobacterium has shown that it is possible to get foreign DNA into plant chromosomes. Can biologists mimic the natural process with a plasmid of their own choice or design, or does the symbiotic combination of *A. tumefaciens* and the *Ti* plasmid have inimitable advantages shaped by evolution and natural selection? Until now counterfeit plasmids introduced into plant cells by a variety of techniques have failed to integrate their DNA into the host DNA. Even the *Ti* plasmid itself nearly fails when it is introduced artificially into tobacco protoplasts (cells denuded of their tough outer wall). The efficiency of *T*-DNA insertion is very low and the ends of the inserted DNA are abnormal, suggesting that the usual integration mechanism is not functioning.

It is a safe bet that, in contrast, *Agrobacterium* knows how to get *T*-DNA into the plant-cell nucleus with high efficiency. How? There may be a bacterial enzyme that excises *T*-DNA from the plasmid. The form of the *T*-DNA on arrival in the nucleus must be important. The DNA could be linear or circular, single-strand or double-strand, naked or complexed with bacterial proteins; it could replicate independently in the plant cell or it could become integrated immediately. The infecting bacterial cell may somehow bring the plant cell into a receptive state for DNA insertion, or the bacterium may have a special affinity for receptive cells. The signal sequences at the end of the *T*-DNA may be important for excision from the plasmid or for insertion into the plant DNA; what structure (and in which organism) reads the signals is not known.

For all its mysteries the *Ti* plasmid is today the only successful gene vector for higher plants. Quite apart from the great promise of its exploitation for crop improvement, the plasmid has become an important investigative tool. It provides a way to manipulate plant genes and thereby to study gene structure and regulation. It has revealed genes that influence plant-hormone levels, and such genes will be valuable for the study of development. Finally, *T*-DNA has demonstrated for the first time that bacterial plasmid DNA recombines with plant chromosomal DNA in nature, thus revealing a previously unknown evolutionary channel for the transmission of genes.

Giant Volcanic Calderas

They are craters tens of kilometers in diameter that remain after eruptions far more violent than any in the span of human history. Perhaps 10 of them have been created over the past million years

by Peter Francis

The eruption of Mount St. Helens in southern Washington on May 18, 1980, ejected .6 cubic kilometer of magma and left a crater two kilometers in diameter; it is deservedly regarded as having been a spectacular event. How then might one describe an eruption that happened 950 kilometers to the east of Mount St. Helens some 600,000 years ago? It disgorged 1,000 cubic kilometers of pumice and ash and left an elongated caldera (the term for a large volcanic crater) 70 kilometers in its longest dimension. The camouflaging effects of vegetation and glaciation have made the traces of that eruption quite difficult to recognize today; the most obvious vestige is the Old Faithful geyser in Yellowstone National Park. Yellowstone is nonetheless a product of volcanic processes operating on the largest scale: it is a resurgent caldera, that is, a caldera whose floor has slowly domed upward in the millenniums since the eruption. Resurgent calderas are by far the largest calderas found on the earth. Thus volcanic eruptions such as the one that formed the Yellowstone caldera must be counted among the greatest catastrophes of nature, perhaps comparable to the impact of an asteroid.

Fortunately they are rare. Not one has taken place throughout the few thousand years of recorded history, and in the U.S. only three are known to have happened over the past million years. In addition to the one in Yellowstone, an eruption 700,000 years ago formed the Long Valley caldera in California and an eruption a million years ago formed the Valles caldera in New Mexico. (The Long Valley caldera has been of particular concern lately because of a series of small earthquakes in the area.) Eruptions of similar age may eventually be shown to have happened in other parts of the world. Still, it will probably turn out that in the past million years no more than 10 have happened worldwide. On the other hand, detailed mapping of the San Juan Mountains in Colorado by Thomas A. Steven and Peter W. Lipman of the U.S. Geological Survey revealed at least 18 calderas between 20 and 30 million years old, and many others of comparable age have been recognized in southern New Mexico, Arizona and Nevada. In the past decade volcanologists have made rapid progress in understanding the origins of giant, resurgent calderas and the catastrophic eruptions that form them.

The fundamental mechanism of caldera formation is clear. The sudden ejection of large volumes of magma from a magma chamber a few kilometers below the surface of the earth abruptly removes the underpinning of the roof of the chamber. Hence the roof founders downward, leaving a caldera at the surface. In a classic work on the volcanic geology of Scotland published in 1909, C. T. Clough, H. B. Maufe and E. B. Bailey offered the first account of this process, which they termed cauldron subsidence. The process takes place on a wide range of scales, leaving calderas that vary in diameter from a few kilometers to 50 or more.

Apart from sheer size the definitive feature of a resurgent caldera is the slow upheaval of its floor, which probably results from the intrusion of fresh magma into the magma chamber that created the caldera in the first place. The amount of vertical resurgence can exceed one kilometer. Unlike an ordinary volcano, therefore, a resurgent caldera is a broad depression with a central massif. The phenomenon of resurgence was first recognized in 1939 by the Dutch geologist R. W. van Bemmelen during a study of the Toba caldera in northern Sumatra. Van Bemmelen estimated that the floor of the caldera had subsided by as much as two kilometers, allowing a deep lake to form, but had later been heaved up hundreds of meters to create the island known as Samosir in the middle of the lake. Toba remains today the largest resurgent caldera known: its maximum dimension is nearly 100 kilometers. Toba, however, has been relatively little investigated, and the details of resurgence have been worked out largely by geologists of the Geological Survey who have studied the calderas of the American Southwest. Indeed, the term resurgent caldera was introduced (in 1962) by Robert L. Smith and R. S. Bailey of the Survey.

One further characteristic of resurgent calderas can be traced to the nature of volcanic processes. In any volcanic eruption the magma reaching the surface can erupt in three ways: as lava, as air-fall material or as a pyroclastic flow. Lava is simply magma that reaches the surface, spreads out as a liquid and solidifies into igneous rock that is finely crystalline or even glassy. Depending on the composition of the magma, the rock ranges from basalt, a dark gray material relatively poor in silica (SiO_2), through andesite and dacite to rhyolite, a light gray material rich in silica. Air-fall material consists of pumice (a frothy, glassy substance) along with finer particles of ash and dust. (The term dust is usually given to particles less than four micrometers in diameter. Particles between four and 63 micrometers are called fine ash.) The material is created when fragments of solidified magma are carried

CERRO GALÁN CALDERA in northwestern Argentina went undetected in conventional geological surveys of the area but was discovered in this false-color image, which was made by a Landsat satellite in orbit 900 kilometers above the ground. The sheer size of the caldera had helped to hide it. The rim of the caldera, in places covered with snow, is an ellipse 34 kilometers along its greatest (north-south) diameter. The center of the caldera, also covered with snow, is an elevation whose apex is six kilometers above sea level. The elevation (a resurgent center) exists because the accumulation of fresh magma under the surface in the millenniums since the catastrophic eruption at the site has heaved the caldera floor upward. The grayish, heavily gullied terrain surrounding the northern half of the caldera is the much-eroded remnant of a great apron of ignimbrite: rock that formed from the pumice and ash ejected in the eruption.

upward in the atmosphere by intensely hot convecting gases.

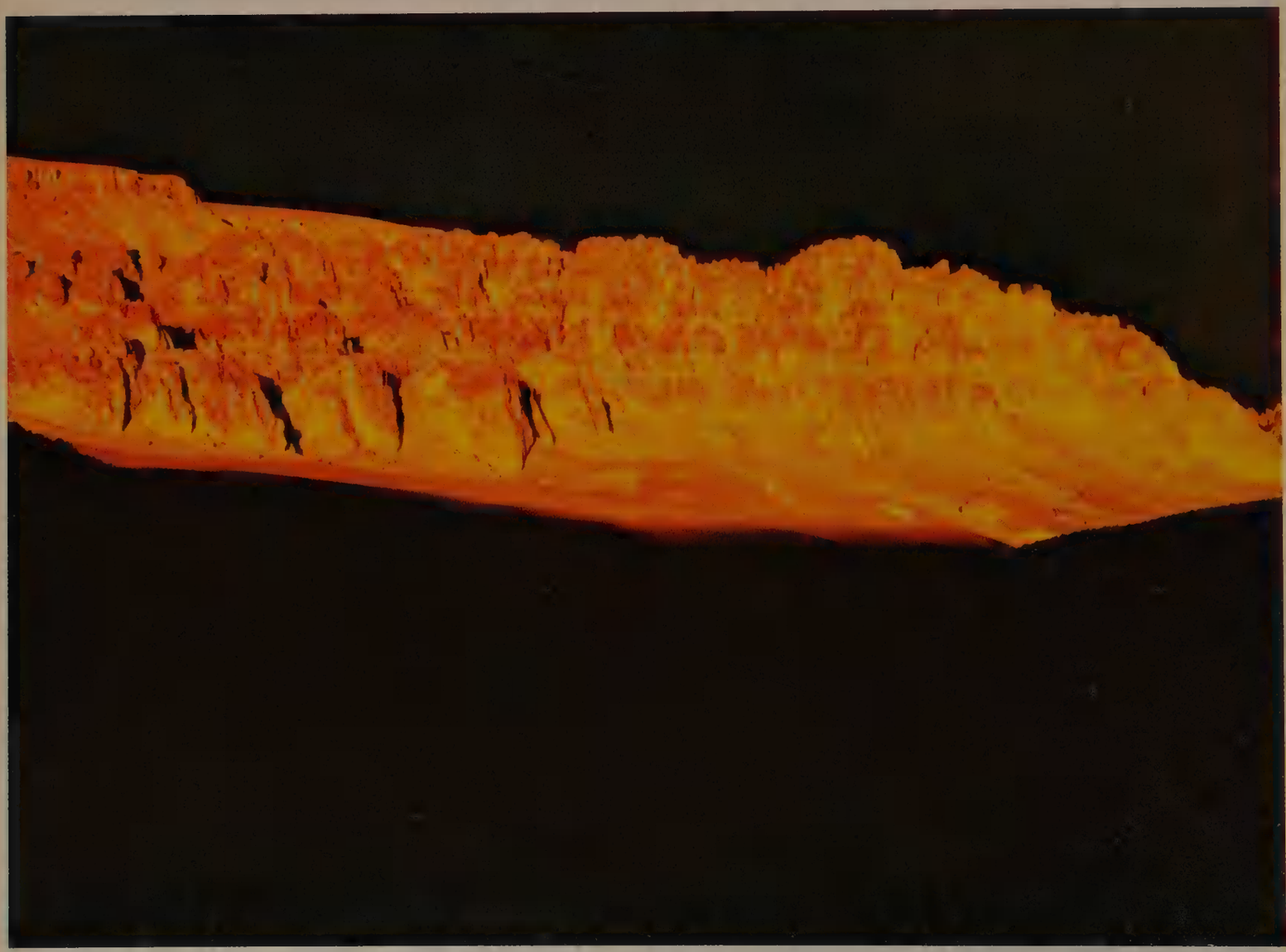
Pyroclastic flows are also composed of pumice, ash and dust. Here, however, the volcanic material forms an incandescent, ground-hugging cloud that races along the surface at speeds of up to 100 meters per second, buoyed up and fluidized by hot gas. Often the cloud deposits clasts, or fragments, of pumice many tens of kilometers from the source of the eruption.

Most volcanic eruptions produce lava and air-fall ashes; pyroclastic flows are less common. Yet in eruptions that form a resurgent caldera pyroclastic flows account for by far the greatest proportion of the ejecta. Indeed, pyroclastic flows often accumulate in a newly formed caldera to a thickness of more than a kilometer. At the bottom of such a pile the pumice clasts typically soften and weld together, producing a distinctive and unmistakable rock. In 1935 P. Marshall, studying the rock resulting from pyroclastic flows in New Zealand, named it ignimbrite: fire-cloud rock. The term is now employed for the deposits laid down by pumiceous pyroclastic flows, whether the clasts are welded or not. Large thicknesses of welded ignimbrite are excellent evidence for the location of ancient calderas.

Resurgent calderas are large and the material ejected by the eruptions that make them is extremely voluminous. Nevertheless, resurgent calderas are not particularly easy to identify. In many instances they are camouflaged by their very size. A few years ago M. C. W. Baker and I, working first at the Open University in England and later at the Lunar and Planetary Institute in Houston, undertook a search for calderas in the Andean cordillera of South America by means of images made with the Landsat satellites. The search capitalized on the greatest advantage that satellite imagery holds over conventional aerial photography: the altitude of the spacecraft (900 kilometers) gives an otherwise unobtainable synoptic view of large structures on the earth. We were able to identify two major calderas that had previously gone undetected by geological surveys.

The more impressive of the two is the Cerro Galán caldera in northwestern Argentina, which is 34 kilometers in diameter and is surrounded by a spectacular apron of ignimbrite extending as far as 70 kilometers from the caldera rim itself. We later learned that the caldera had also been noted by J. D. Friedman of the Geological Survey and Grant H. Heiken of the Johnson Space Center in photographs made from the *Skylab 4* spacecraft in 1974. The caldera is so young (2.6 million years) and its structure is so unambiguous that we had little doubt what it is.

The second caldera was less obvious. The Landsat images we examined suggested to us that the Kari Kari mountain massif in Bolivia, which is five kilometers high, probably represents the resurgent center of a large, old caldera. Earlier mapping of Kari Kari, however, had suggested the massif was a batholith: a great mass of coarsely crystalline igneous rock that solidified in the crust of the earth and was later exposed by erosion. Our field work at the site soon confirmed our identification. The welded texture of the rock of the massif showed conclusively that it is ignimbrite. The resurgent



CLIFF OF IGNIMBRITE nearly half a kilometer high was exposed by erosion at Cerro Galán at a site on the apron of ignimbrite about 20 kilometers to the west of the caldera rim. Each of the horizontal layers in the cliff represents material deposited by an individual eruption; the uppermost layer represents material deposited by the eruption that formed the Cerro Galán caldera some 2.6 million years ago.

center is the remaining evidence of a caldera that was originally about 36 kilometers across in its largest dimension. It is 20 million years old. We have found a number of other calderas in unexplored parts of the central Andes. Field investigation will be long and difficult.

At first one might think the most likely place for an eruption that forms a resurgent caldera would be a subduction zone, the boundary on the surface of the earth where a plate of oceanic crust slides under a continental plate and plunges into the underlying mantle. After all, subduction zones are sites of intense seismic and volcanic activity. The Toba caldera in Sumatra is in such a setting. In most instances, however, the geological setting is much less straightforward. Most of the younger calderas in the U.S., for example, are hundreds of kilometers from any modern subduction zone.

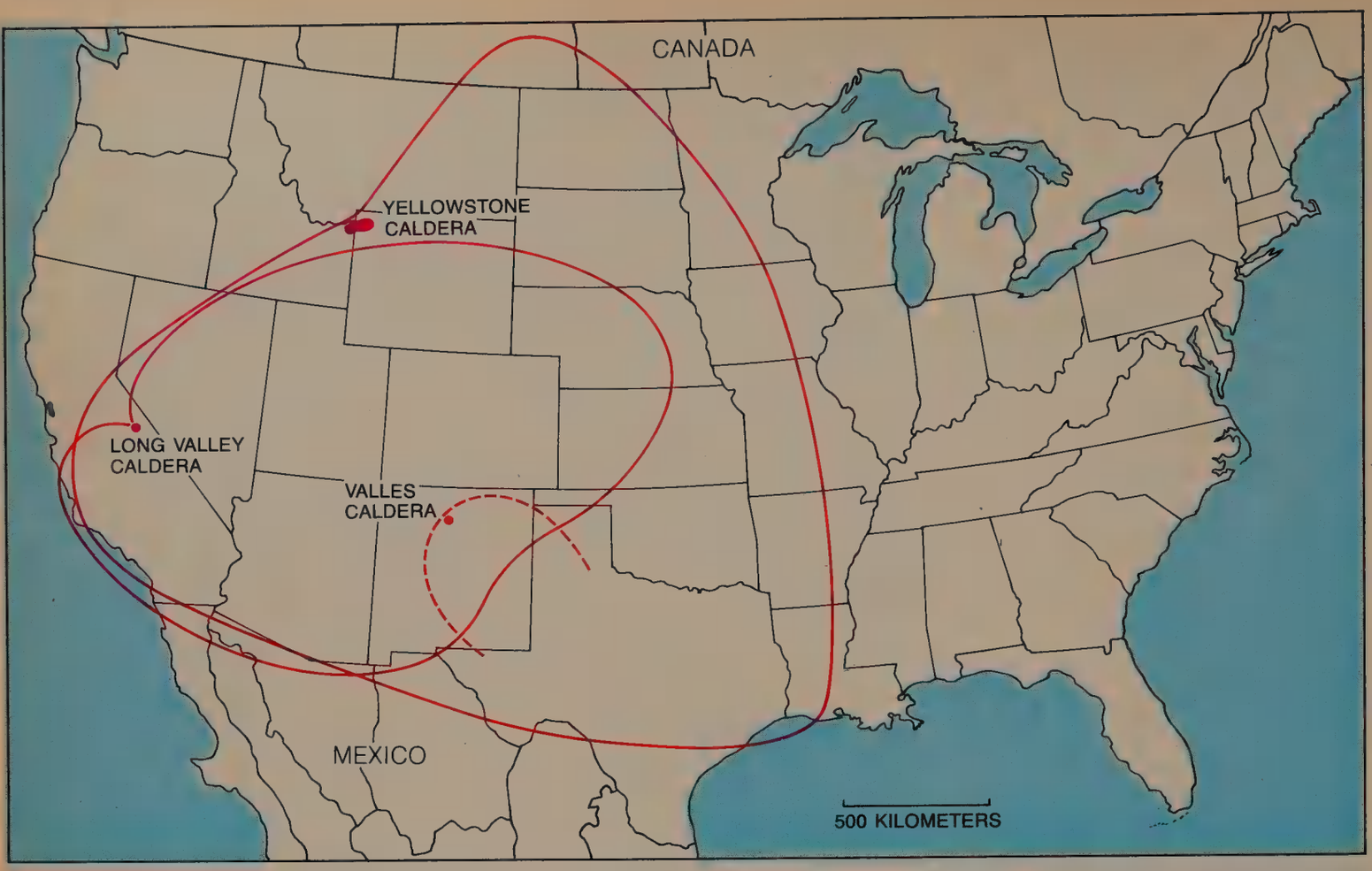
Nevertheless, resurgent calderas are not randomly distributed around the globe. The ignimbrites that characterize them result from the eruption of dacitic or rhyolitic magma, which is viscous, rich in silica and typically is produced in areas where the continental crust is thick. Resurgent calderas may form, therefore, in regions of the continental crust where a thermal plume (a "hot spot") in the mantle of the earth is large enough and long-lasting enough to melt huge volumes of rock. The plume does not melt the continental crust directly. Instead it melts part of the mantle to create a basaltic magma. The basaltic magma rises, melting the rock at shallower levels.

Consider some specific calderas. In the U.S. the Yellowstone caldera lies at the northeastern end of a trail of volcanic activity that begins in Idaho in the basaltic rock of the Snake River plain. Over the past 15 million years the focus of volcanic activity has shifted along the trail to its present position in Wyoming, perhaps in response to the movement of the plate that includes the North American continental crust over a stationary thermal plume in the mantle. A number of other calderas, no older than a few tens of millions of years, are found in a zone many hundreds of kilometers wide in Nevada, Arizona, Utah and New Mexico. The youngest caldera in the group is on the flanks of the Rio Grande Rift, which runs for hundreds of kilometers northward through New Mexico into Colorado. It is thought the continental crust at the Rio Grande Rift has somehow been thinned, producing the rift itself. A similar process is thought to have caused rifts in the oceanic crust near many of the island arcs in the Pacific. At the Rio Grande Rift the thinning may have allowed heat from the mantle to act with great effect.

Other settings are also possible. In Ar-

WELDED IGNIMBRITE, shown actual size, was photographed on the flanks of the Valles caldera in New Mexico. The dark streaks are compressed clasts, or fragments, of glassy pumice (a frothy volcanic rock) that softened and fused in the heat of the layer of ejecta laid down by the caldera-forming eruption. The clasts now lie in the midst of a hard, dense rock.

UNWELDED IGNIMBRITE was photographed on the flanks of the Cerro Galán caldera. The clasts of pumice lie in a matrix of volcanic ash and dust. They can be removed by hand.



RAIN OF ASH is mapped for the three eruptions that left resurgent calderas in the U.S. over the past million years. An eruption a million years ago left the Valles caldera in New Mexico, an eruption 700,000 years ago left the Long Valley caldera in California and an eruption 600,000 years ago left the Yellowstone caldera in Wyoming. The mapping was done largely by G. A. Izett and his colleagues in the U.S. Geological Survey; the uncertain extent of the Valles caldera's ash is shown as a broken line. Ordinary volcanoes can distribute ash over large areas, but the thickness of the ash layer laid down by a caldera-forming eruption is centimeters rather than a fraction of a millimeter.

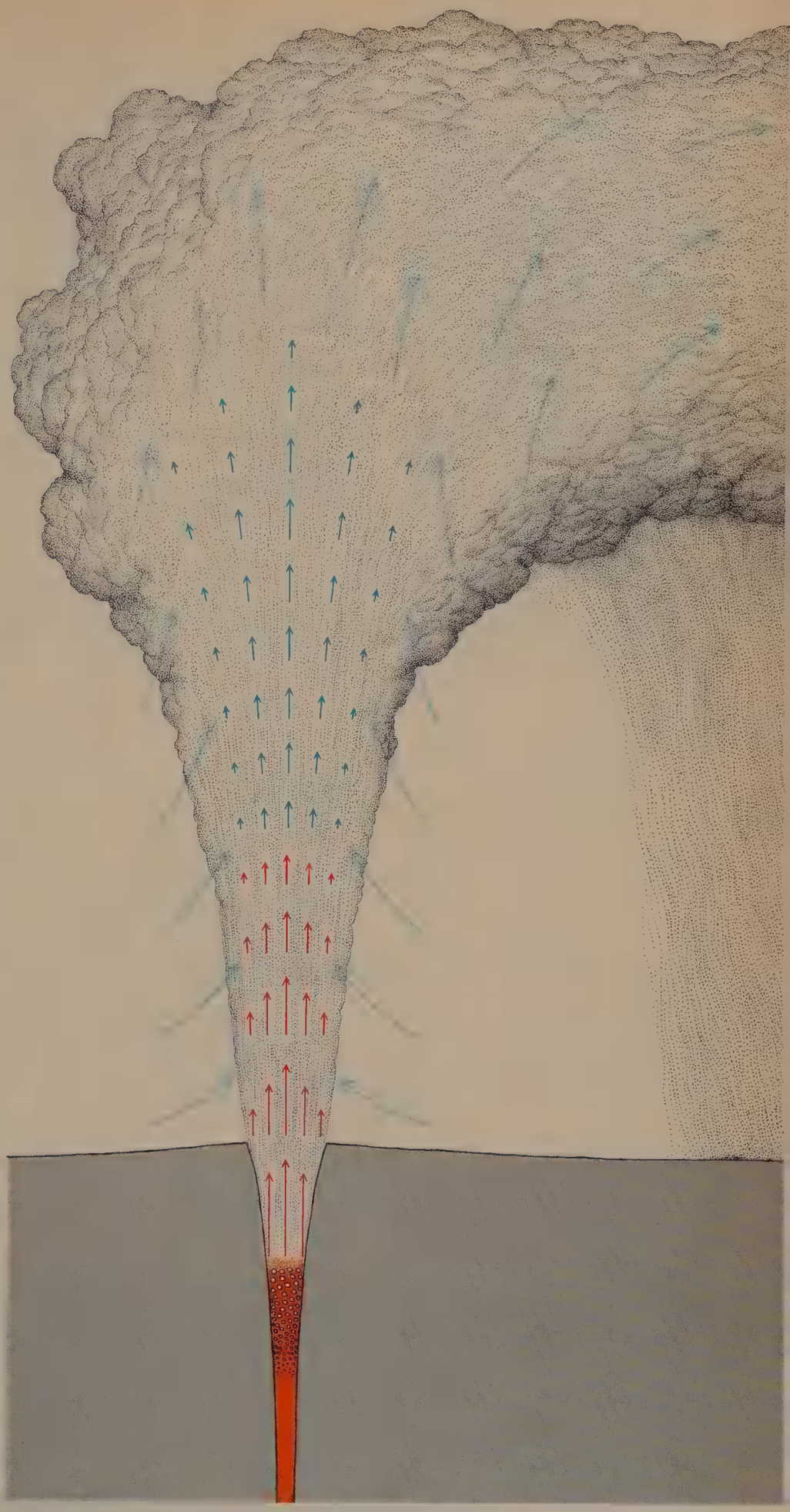
ASH FROM THE TOBA CALDERA in Sumatra rained down on a region including most of the Bay of Bengal and parts of Sumatra, Sri Lanka and India 75,000 years ago. The pattern of distribution was determined by a group led by Dragoslav Ninkovich of the Lamont-Doherty Geological Observatory of Columbia University; the group found an ash layer more than 10 centimeters thick in cores extracted from the sea floor as far as 2,000 kilometers from the Toba caldera itself. The Toba caldera is the youngest and the largest caldera known.

gentina and Bolivia resurgent calderas have formed not only along the main volcanic cordillera of the Andes but also in a second cordillera more than 200 kilometers farther inland. Here one finds no clear evidence of crustal thinning. On the contrary, the continental crust may be as much as 40 to 50 kilometers thick under the Cerro Galán and Kari Kari calderas. It is thought magma conduits to the surface of the inland cordillera may have developed by local extensional stress: a fracturing of the crust caused by the pressure of magma moving upward because of its buoyancy.

A number of events can be identified in the evolution of a typical caldera: precaldera doming, caldera collapse, eruption of air-fall material and pyroclastic flows, postcaldera resurgence and finally late-stage extrusions of lava. Precaldera doming is the elevation of the surface of the earth that precedes a massive eruption. It happens when a great volume of magma intrudes itself into a shallow level of the continental crust, creating a pluton, or magma chamber, whose top may lie only four or five kilometers below the surface. The doming generates a local extensional-stress field in the surface rock, and the stresses are important to the next event in the evolution: the collapse of the caldera.

An ambiguity persists, however. Does the upward pressure of the pluton cause the failure of the rock that forms the roof of the chamber, allowing magma to escape to the surface along a great ring-shaped fracture that develops above the pluton's perimeter? The escape of the magma would open a subterranean space into which the roof could then subside. Or does the roof of the pluton sink into the magma and in the process open the ring fracture along which magma erupts? T. H. Druitt and R. S. J. Sparks of the University of Cambridge have argued that whether the subsidence is a cause or an effect of the eruption, the ring fracture, which becomes the wall of the caldera, must be vertical or must bell outward with increasing depth in order to accommodate the rate of eruption and the pistonlike fall of the roof.

The magma at the top of the pluton has a temperature of 700 to 1,000 degrees Celsius and is rich in dissolved gases, mostly water vapor. The magma rises toward the surface along the newly opened ring fracture. As it rises, the pressure on it lessens, until at a depth of about a kilometer the gases come out of solution, much as they do when the cork on a bottle of champagne is popped. Dacitic or rhyolitic magma is much more viscous than champagne, however (or even basaltic magma), and so the gases do not merely bubble away. Instead they blow the magma apart. The actual proc-



PLINIAN COLUMN is generated in ■ volcanic eruption when the rate of eruption is great, when the erupting magma includes large amounts of dissolved gas and when the fissure venting the magma is narrow: perhaps 50 to 100 meters in diameter. At a depth of about a kilometer the gas comes out of solution, causing the magma to explode into incandescent pumice and ash. The particles are blasted upward ballistically (*solid red arrows*) to form the lower part of the plinian column. In the upper part of the column they are lofted many kilometers higher (*solid blue arrows*) by convection currents set up in the atmosphere. It is thought that plinian columns are often the first phase in the catastrophic eruption that forms ■ resurgent caldera.

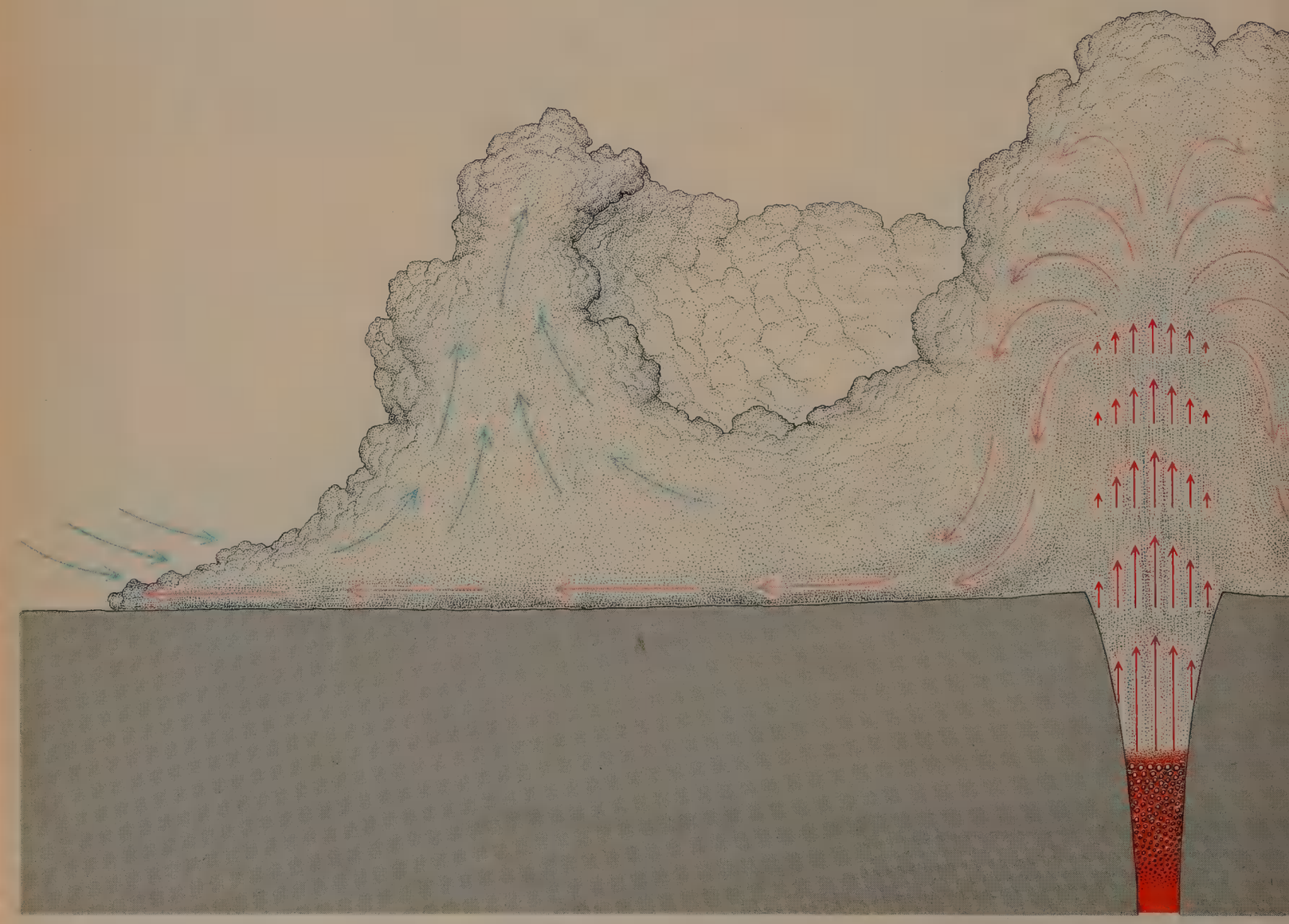
ess is complex but the result is simple: the magma escaping from the pluton toward the surface is expanded into pumice and explosively fragmented into incandescent solid particles ranging in size from micrometers to meters.

The key to understanding the effects of a great caldera-forming eruption is the action of the incandescent mixture of dust, ashes, pumice and liberated gases when it is vented at the surface. The kinetics of the venting have been analyzed by a group originally centered on G. P. L. Walker of the Imperial College of Science and Technology, notably Sparks of Cambridge, Lionel Wilson of the University of Lancaster and Colin J. N. Wilson of the University of Auckland. They have shown that when the rate of eruption is great and the vent is relatively small (perhaps 50 or 100 meters in diameter), an eruption column develops, rising tens of kilometers into the atmosphere. (The eruption column of Mount St. Helens on May 18, 1980, reached about 20 kilometers.)

It should be emphasized that the pumice in the column is not simply propelled upward from the vent like buckshot from a shotgun. Directly above the vent the energy of the column indeed is largely kinetic. The upward velocities are hundreds of meters per second. As the pumice rises, however, it rapidly decelerates; it is slowed not only by gravity but also by aerodynamic drag. Then a second process begins to contribute energy. The decelerating assemblage of incandescent pumice, ash and gas entrains and heats up air from around the column. As a result the assemblage becomes buoyant and begins to rise convectively. It may even accelerate upward again. Eruption columns rising by convection are well known; they are the essence of what is called a plinian eruption, after Pliny the Elder, whose description of Vesuvius in A.D. 79 provides the first documented example. Convection can drive a plinian column to heights as great as 50 kilometers.

Massive plinian columns may signal the start of a catastrophic collapse that creates a caldera. As the eruption proceeds, however, the plinian columns typically give way to pyroclastic flows, which account for by far the greatest fraction of the volume of the eruption. The reasons are several. The size of the vent can be increased by the blast of the initial eruption or by the collapse of the caldera floor. The gas content of the magma can decrease as lower levels of the pluton are tapped. In these circumstances it becomes more difficult for a plinian column to sustain itself. The column becomes denser than the surrounding air and soon collapses. That is, the mass of incandescent pumice and gas acquires only the first, ballistic component of upward motion. It still may reach a height of 10 kilometers, but without convective flow to sustain it, it falls back around the vent.

Now the pyroclastic flows develop. They are made up of the falling material, and they surge outward from where the material reaches the ground. Their mobility is extraordinary. It is known from the distribution of the ignimbrites they deposit that they can sweep over hills as much as a kilometer high and travel distances of up to 150 kilometers.



PYROCLASTIC FLOW is generated in a volcanic eruption when the rate of eruption slows, when the erupting magma includes little gas or when the vent widens. Often such a flow results when a plinian column collapses, so that volcanic particles rise only ballistically (*solid*

No large-scale pyroclastic flows have ever been observed in action, but their hill-climbing abilities imply a velocity on the order of 100 meters per second. The smaller flows in the later stages of the eruption of Mount St. Helens had velocities of 30 meters per second.

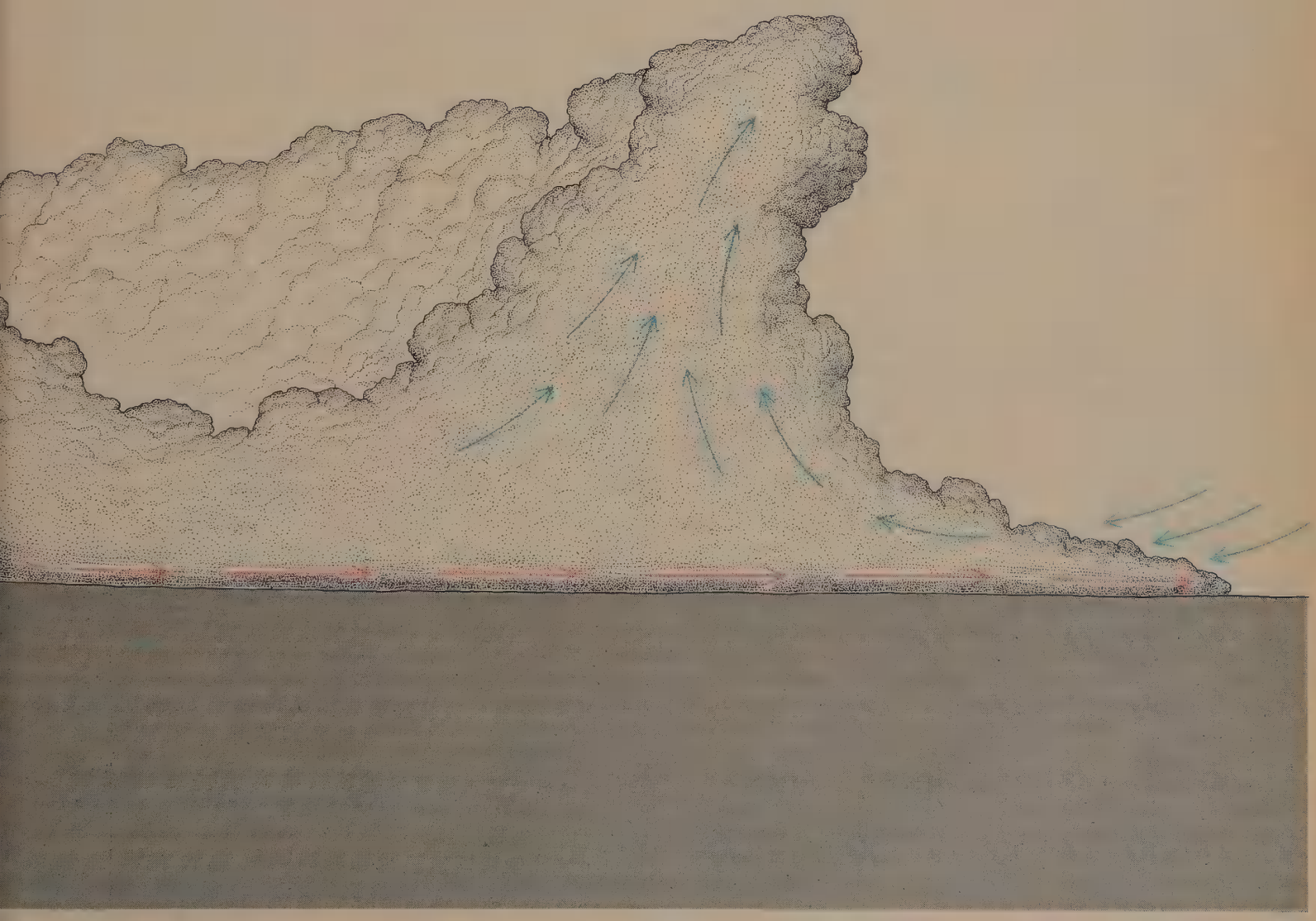
The explanation of the mobility is fairly well established as a result of recent theoretical and experimental studies such as those of Sparks and Colin Wilson. The flows travel in a partially fluidized condition. That is, the gases that continue to leave the hot pumice particles, combined with the atmospheric gas entrained at the front of the flow, stream upward among the particles, supporting the finer grains. The finer grains in turn provide a lubricating medium in which the larger clasts are dispersed. The flow thus comes to have a lower density and a much lower viscosity than an unfluidized aggregation of pumice and dust would have. It makes efficient use of its kinetic energy, which derives from its vertical fall from a height of several kilometers.

The main deposit of a massive pyroclastic flow is a blanket of pumice and smaller particles that can be many meters thick more than 50 kilometers from the vent. In addition the fine particles entrained by the flow typically form a secondary ash cloud that rises many kilometers by convection. The subsequent fall of particles from the cloud can deposit a thin layer of ash over a region much larger than the one covered by the ignimbrite of the pyroclastic flow itself. Indeed, the layer, which is called co-ignimbrite ash, can amount to as much as a third of the entire volume of the ignimbrite.

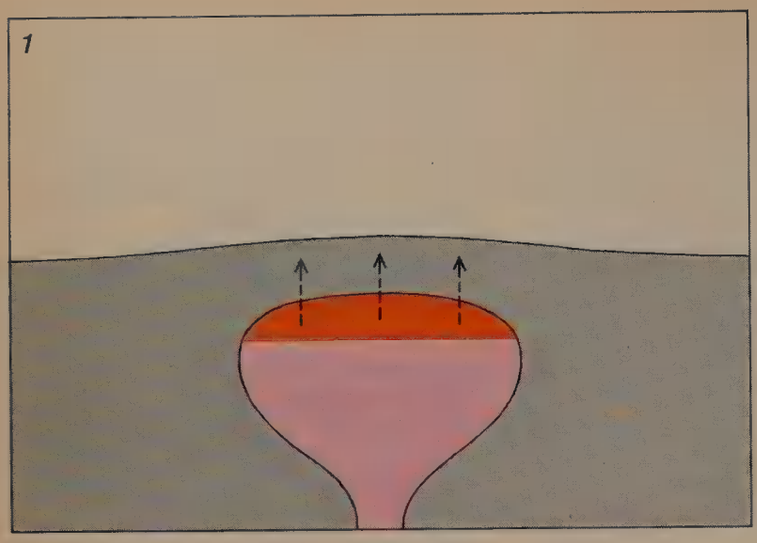
Why are the eruptions that form resurgent calderas exclusively eruptions of dacitic or rhyolitic magma? Two interrelated factors are responsible. First, because of the lower viscosity of a basaltic magma, gases coming out of solution in the magma can easily escape. The eruption is not catastrophically explosive, and the production of pumice and fine ash is inhibited. Second, because a basaltic eruption gives rise to scant fine ash, the transfer of heat from the eruption column to the atmosphere is lessened. (Such transfer is far more efficient for particles less than a millimeter in diameter than it is for larger ones.) Hence it is difficult for an eruption of basaltic magma to generate convective currents in the atmosphere and become a plinian column.

The "fire fountains" seen on the active volcanoes of the Hawaiian Islands provide an excellent example. In a fire fountain great volumes of lava are sprayed high into the air, but the lava, which is basaltic, emerges as large liquid gobbets, sometimes a meter or more across. Hardly any fine ash erupts. The gobbets do not readily transfer heat to the atmosphere, and so there is virtually no chance that a convective column might develop. Instead the eruption leaves lava spattered on the ground around the vent. Since the lava retains almost all its original heat, secondary lava flows are common.

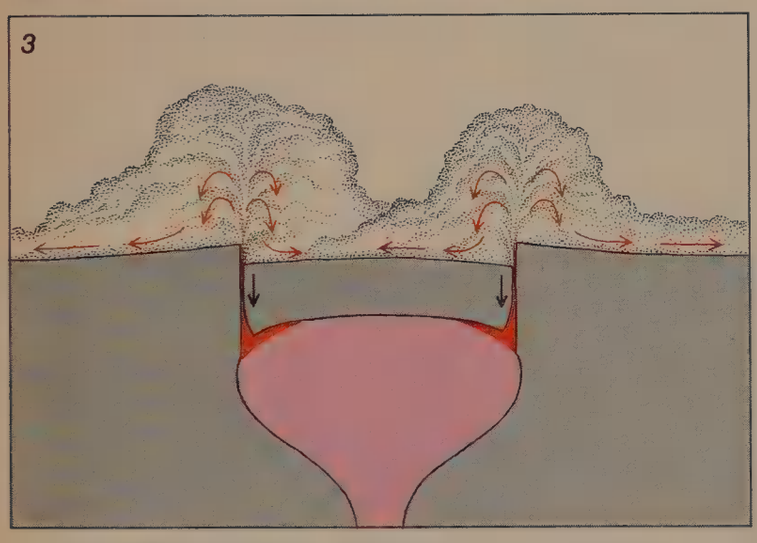
One of the most surprising aspects of the catastrophic eruptions that form a resurgent caldera is how short they are. Michael T. Ledbetter of the

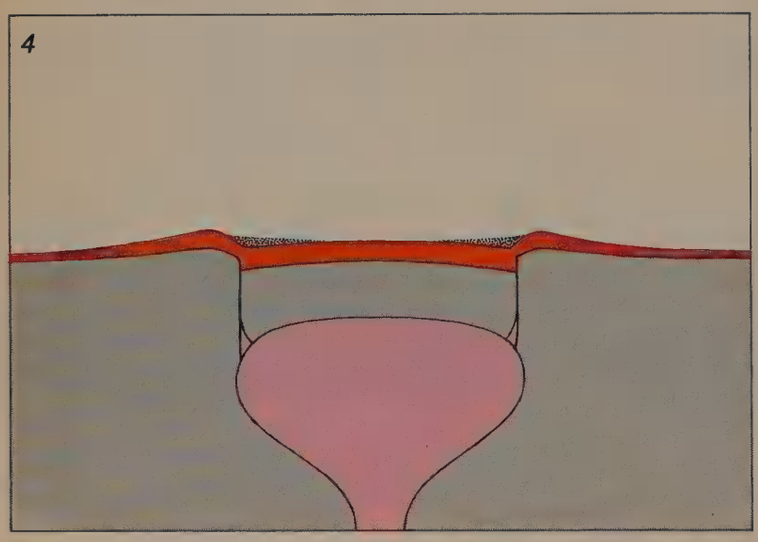


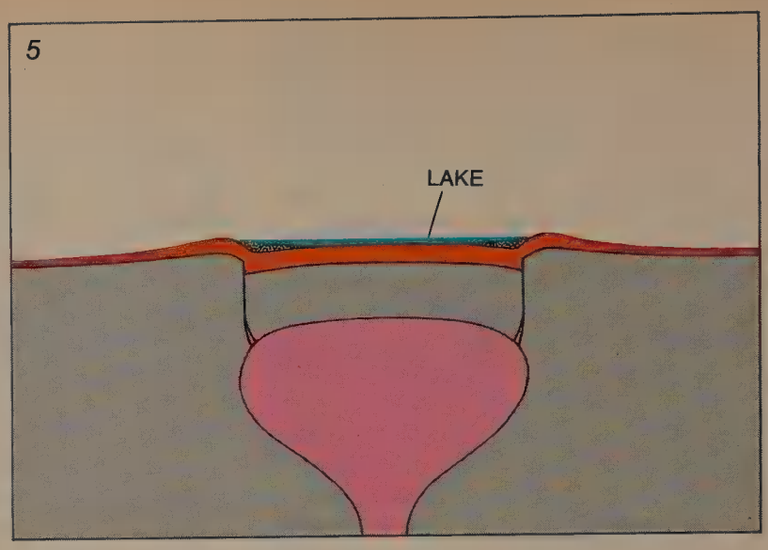
***red arrows*) and then fall back to the ground. There they form the flow: a ground-hugging cloud of particles buoyed up and fluidized by hot gases. The flow can climb hills a kilometer high and can travel distances of 150 kilometers at velocities of 100 meters per second.**

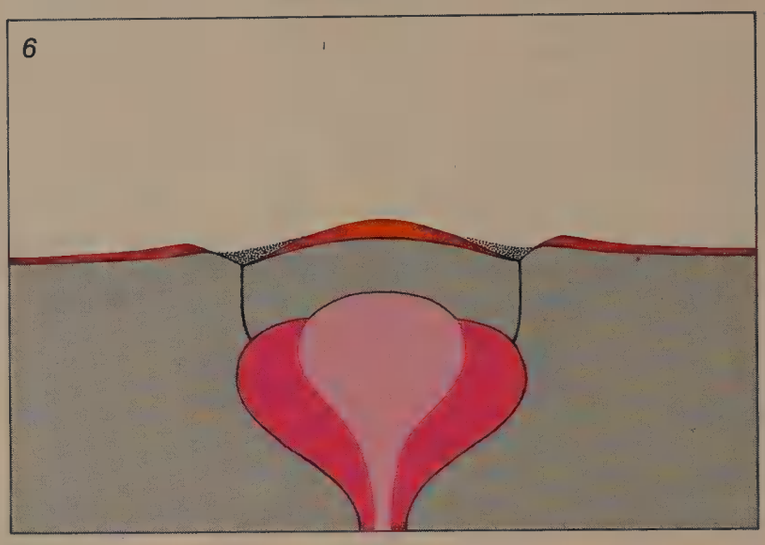

1

2

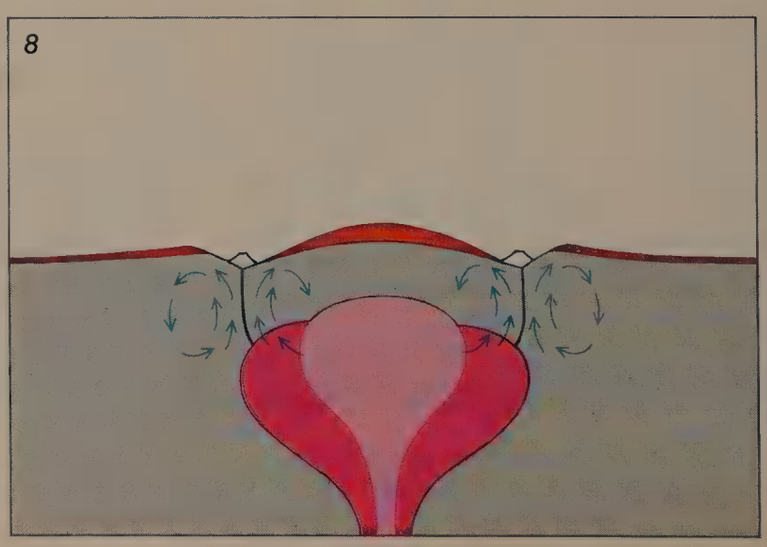

3


4


5
LAKE


6

7


8

University of Georgia and Sparks have studied the rate at which ash particles settle in seawater. The rate depends on the density and the radius of each particle; therefore fine ash among the ejecta from the beginning of an eruption and coarse ash among the ejecta from the end of the eruption may reach the sea floor simultaneously. Ledbetter and Sparks deduce from the data they have collected that the 300 cubic kilometers of ash erupted from the Atitlán caldera in Guatemala 84,000 years ago fell over a period of only 20 to 27 days. They think it unlikely that the eruption was continuous. Instead a plinian column probably collapsed a number of times, causing the successive emplacement of distinct ignimbrite sheets. (The Atitlán caldera is 28 kilometers across, but its center is not resurgent.) The same approach suggests the eruption that formed the Toba caldera deposited more than 1,000 cubic kilometers of ash in as little as nine days.

After the eruption several things happen. Typically a lake fills the new caldera. Sediment erodes from the caldera wall and accumulates at the bottom of the lake. On a slower scale the resurgence of the caldera floor begins. In general it is not a simple pistonlike rise of the floor (and of the ignimbrite that now blankets it) along the ring fracture. Mapping of the ignimbrite in Cerro Galán and other calderas shows that it is tilted away from a well-defined central point, often at an angle exceeding 45 degrees. This suggests resurgence is a local phenomenon. Probably it occurs above a small pluton newly intruded into the one that caused the eruption. In some cases, including Yellowstone, two separate centers of resurgence are present in one caldera.

It was the recognition that young lake sediments had been raised hundreds of meters that enabled van Bemmelen to demonstrate that the resurgence of the floor of the Toba caldera had formed the 640-square-kilometer island of Samosir. At Cerro Galán resurgence of more than one kilometer has raised the center of the caldera to an altitude of more than six kilometers above sea level, making it one of the highest mountains in Argentina. Little is known about the rate of resurgence, but in the case of the Long Valley caldera, the one at which the dating of the geological events is least equivocal, the process may have continued for up to 200,000 years. Since the Toba caldera is only about 75,000 years old, the resurgence of Samosir may not yet be complete.

After resurgence come the final episodes in the evolution of a caldera: the relatively quiet effusion of dacitic or rhyolitic lava from a necklace of vents along the ring fracture. Typically the volume of material released is small, but the effusions continue intermittently long after the catastrophic caldera-forming eruption. At Long Valley distinct episodes of effusion took place 500,000, 300,000 and 100,000 years ago. It is fair to conclude that the volcanic events associated with the formation of a caldera may continue with little violence for up to a million years. Hot springs and geysers, representing geothermally heated water that finds its way to the surface, may be present for much longer than that.

The hazards to human life that would be presented by a caldera-forming eruption have never been analyzed in detail. They would surely be far-reaching. Much would depend on the total volume of the eruption and on the proportions of air-fall ash, ignimbrite and co-ignimbrite ash. Consider a 1,000-cubic-kilometer eruption. A tract of land with a surface area of perhaps 500 square kilometers would sink, and the resulting caldera would be filled entirely with ignimbrite. A surrounding area of up to 30,000 square kilometers would also be covered by ignimbrite. The depth of the cover would range from more than 100 meters on the caldera rim to a few meters at the farthest extent of the ignimbrite. Throughout the covered area plants and animals would be killed. Buildings would be demolished or buried.

In the short term the most serious consequence of the eruption would probably be the fall of co-ignimbrite ash, which would have a pattern dictated by the prevailing wind and the size distribution of the particles. Particles much larger than about 80 micrometers stay in the atmosphere for a few hours at most. Smaller particles stay aloft for much longer. Particles less than 10 micrometers in diameter have settling velocities in the atmosphere no greater than the velocities typical of atmospheric turbulence. They may thus stay airborne for times of up to 100 days and so can be transported great distances. Co-ignimbrite ashes tend to be extremely fine; they would probably be dispersed over several million square kilometers.

Imagine the eruption of Mount St. Helens multiplied more than a hundredfold and taking place at Yellowstone today. Apart from the immediate effects, which would include the immobilization of ground and air traffic by co-ignimbrite ash, the most severe effect would probably be on agriculture, since millions of square kilometers of grain-growing farmland would be covered by ash to a depth of several centimeters. The result is easy to forecast: the disruption of at least one year's harvest, and possibly several. Over a longer term the climate of the earth would be affected by the injection of dust into the atmosphere together with an extremely long-lasting aerosol of particles that arise from acidic volcanic gases. The resulting worldwide disruption of agriculture could persist for a number of years.

Could such an eruption be predicted? The most likely indication of a future caldera-forming eruption might be the "leakage" of dacitic or rhyolitic magma to the surface as a ring fracture develops. In addition seismic signals might indicate the movement of magma into a chamber a few kilometers below the surface. (Of the two types of seismic waves, *S* and *P*, the *S* waves are the most revealing, because they cannot pass through a liquid.) The site of the chamber would be confirmed by a local anomaly in the earth's gravitational field. (The density of magma is less than the density of solid crustal rock.) A further important indicator would be the tumescence, or elevation, of the terrain, and particularly an increase in the rate of tumescence. Conventional surveying techniques can readily reveal such changes.

The difficulty lies in interpreting those various signs as the precursors of a particular event and in placing them on a timetable in the absence of a known case history. Seismic activity and tumescence have recently been noted at the Long Valley caldera and have been taken to indicate the emplacement of fresh magma at a depth of several kilometers. Will a large eruption follow? Or do the events presage merely the quiet extrusion of a new lava dome at the surface? The immediate precursors of a caldera-forming eruption are largely unknown. Moreover, the precursor activity may

DISTINCT STAGES mark the evolution of a resurgent caldera. In the first stage magma at depths of as little as four or five kilometers below the surface forms a pluton, or magma chamber, which slightly domes the surface (*1*). In the pluton there are zones that differ in composition, with viscous magma rich in silica (SiO_2) and dissolved gases at the top (*orange*). Eventually an eruption begins. Plinian columns develop above a great ring-shaped fracture (*2*). Then within minutes or hours the roof of the pluton collapses along the fracture, leaving a caldera, and the plinian columns give way to pyroclastic flows (*3*). The eruption lasts for only a few days. It fills the caldera with ignimbrite and also blankets the surrounding area (*4*). Meanwhile the caldera wall begins to erode (*black stipple*). A lake may also form (*5*). Still later, over a period of a few hundred thousand years, the intrusion of fresh magma into the pluton causes parts of the caldera floor to move upward again (*6*). Minor volcanic activity persists along the ring fracture (*7*). The heat of the pluton may power convection currents of water rich in minerals (*8*) and so give rise to hot springs and geysers at the surface for millions of years.

continue on a time scale spanning many human generations. Hence it may not be possible to predict an eruption correctly. All the same, signs of renewed activity at any known caldera should not be ignored.

The eruptions that form resurgent calderas are so rare that on balance they are probably advantageous to man. A caldera overlying a pluton marks a thermal anomaly in the crust that may continue to manifest itself in hot springs for two or even three million years after the eruption. Apart from the splendid spectacle afforded by the geysers, the hot water and steam are potential sources of useful energy, and so is the hot rock overlying the pluton itself. An important experiment under way in the U.S. is the Hot Dry Rock Project of the Los Alamos National Laboratory. On the flanks of the Valles caldera water is being pumped down a deep borehole into a zone of hot, fractured rock. Heated water returns to the surface through a second borehole. The hot water can be employed for the generation of electricity.

Some less apparent benefits of calderas lie in the useful minerals that are deposited in resurgent calderas or around them over a period of time. The Kari Kari caldera provides an example that is geologically typical and also rich in historical associations. On the flanks of the caldera are the city of Potosí and also the Cerro Rico, a peak whose summit is 4,900 meters above sea level. The peak is composed of rhyolite that was intruded into the caldera's ring fracture about 13 million years ago. As it cooled it was chemically reworked by hydrothermal solutions: hot, acidic liquids ascending after the reaction of water with deep-lying hot rock. In the 17th century Potosí was the largest city in the Western Hemisphere, solely because of the wealth of silver mined from the veins laid down by the solutions. The importance of the mining to Spain is hard to overstate: it gave rise to a well-known Spanish expression, "worth a potosí." Ultimately the flood of silver from Potosí and other mines in the New World disastrously inflated the Spanish currency. Mining has been continuous at Cerro Rico since 1544, but the silver easily won from the upper levels of the deposit has long been exhausted. Tin is now sought at deeper levels.

Clearly the deep-seated sources of heat that create calderas are capable of driving hydrothermal circulation over long periods, and clearly the hydrothermal activity can work complex chemical reactions that allow mineral deposits to precipitate. It may be a while before the details are understood, but the association of calderas with geothermal energy and with valuable mineral deposits ought to inspire gratitude. Without this association resurgent calderas might appear to be unalloyed catastrophes.

TWO LAYERS OF ASH at Crater Lake in Oregon result from an eruption 7,000 years ago that tapped the viscous magma at the top of a pluton and then the less viscous magma under it. The layers of ash reverse that relation. The lower layer, light in color, was deposited first. It is rich in silica and poor in iron and magnesium. The upper, darker layer was deposited immediately afterward. It is poorer in silica and richer in iron and magnesium. Crater Lake's diameter of only nine kilometers makes it too small to be considered a major caldera.

Reshaping the future of film.

From fat to flat. The goal many of us strive for with laborious push-ups and strenuous diets, scientists have achieved in Kodak laboratories. The "bodies" they have streamlined, however, are those of silver halide grains. More than a beauty treatment, this process of reshaping enabled us to position grains for maximum light-catching ability, and expand their sensitizing characteristics. It's this advance in silver halide emulsion technology which made it possible for us to produce a true 1000-speed film with a balance of speed, grain, and sharpness previously unattainable.

The structure of new Kodacolor VR 1000 film is a modification of the inverted-layer format first used in Kodacolor 400 film. This concept places the fastest light-sensitive layers closer to the camera lens. The magenta and cyan layers are repositioned to form a fast-magenta, fast-cyan, slow-magenta, slow-cyan sequence. In the new emulsion, the yellow-filter layer is eliminated, and the protective overcoat and ultraviolet-filter layers are combined into one, without any loss in color quality.

The resultant faster film can be exposed by daylight, electronic flash, or tungsten, which provides photographers with greater picture-taking capability, particularly when recording fast-action scenes or shooting in low-light conditions.

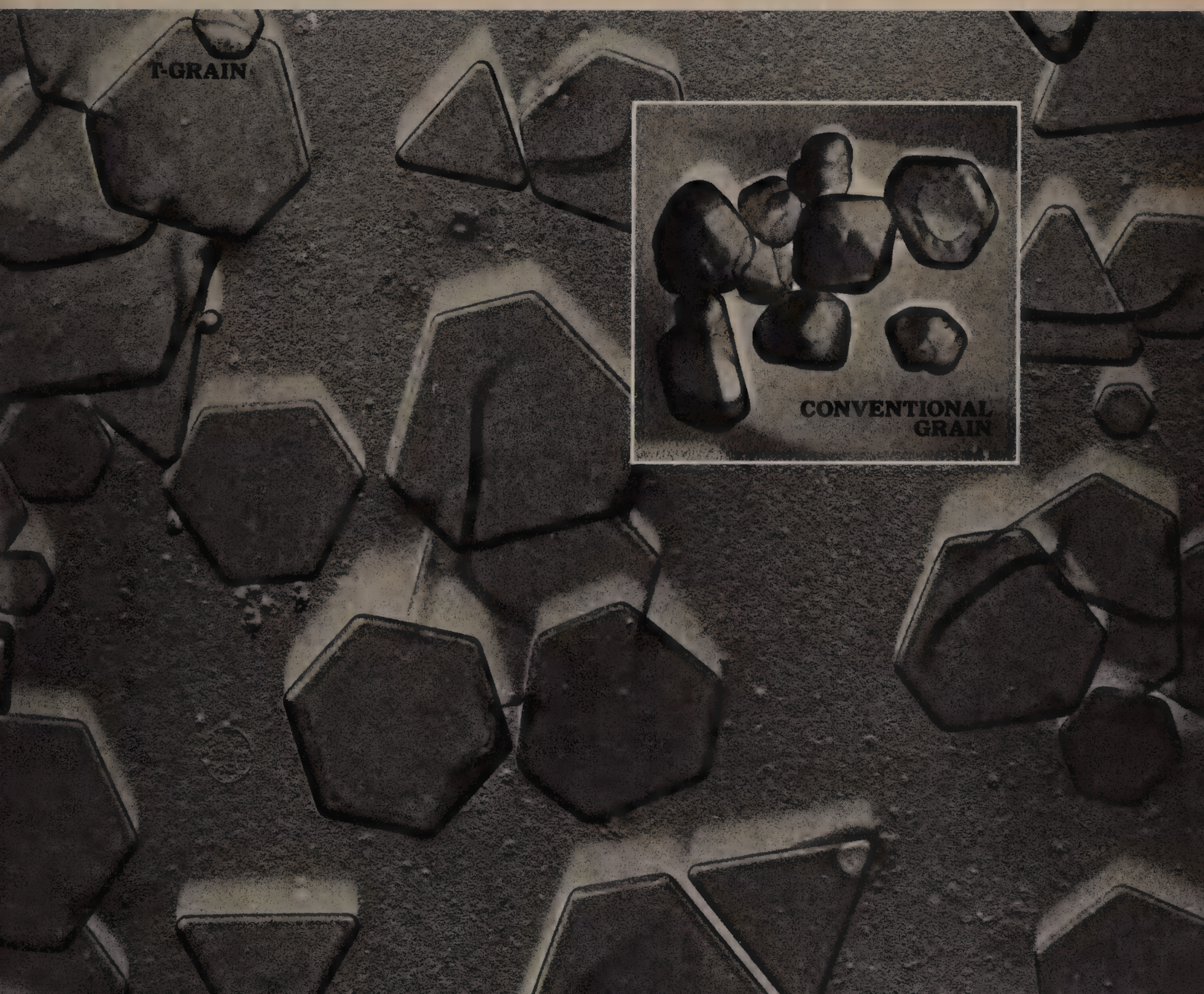
While reshaping the light-sensitive substance of film is a significant accomplishment in itself, we view this new technology as a stepping-stone. It may well reshape the future of film and enable photography to do hundreds of different jobs better than before.

For more information on Kodak's fastest color print film ever and the new emulsion technology behind it...send your request to: Eastman Kodak Company, Department GBSA-8, 343 State St., Rochester, NY 14650.

© Eastman Kodak Company, 1983

Kodak

Kodak. Where technology anticipates need.



Physician, did you miss any of these significant developments in medical science?

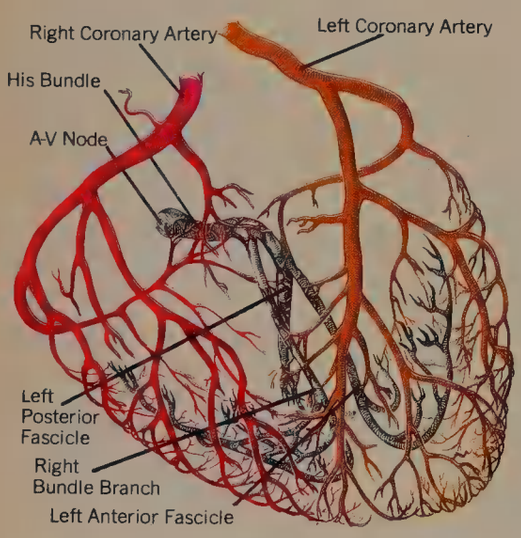
- ***Campylobacter fetus* subsp. *jejuni* is associated with a colitis that can clinically and sigmoidoscopically resemble acute idiopathic ulcerative colitis. Stool cultures are in order for *C. fetus* before beginning nonspecific anti-inflammatory therapy.**

- **Coumarin derivatives cross the placenta. A recent study shows that the consequences for the fetus can be severe. These include embryopathy, stillbirth, and premature delivery.**

- **Nonsteroidal anti-inflammatory drugs may produce a marked reduction in glomerular filtration rate; with termination of the drug, GFR returns to normal.**

- **The first documented incident of indigenous transmission of dengue in the continental United States since 1945 has been reported in Brownsville, Texas.**

IF THESE ITEMS are familiar you must be a prodigiously energetic or prodigiously lucky reader. With 2,000 or more journals published each year, information that significantly affects patient management all too easily slips by. Textbooks are out-of-date before they are published.



Branches of the right and left coronary arteries supply blood to the A-V node and intraventricular conduction system.

SCIENTIFIC AMERICAN *Medicine* is lucidly illustrated with drawings and photographs. Some examples are seen here and on the facing page.

SCIENTIFIC AMERICAN *Medicine* is the busy clinician's answer to this problem.

Because its authors update SCIENTIFIC AMERICAN *Medicine* every month, it is always current. Because the new information appears in a single source, it is there when you need it.

This 2,500-page, innovative union of publishing and electronic technology is the work of leading scholar-practitioners from Harvard and Stanford. The editors are Edward Rubenstein, M.D., F.A.C.P., and Daniel Federman, M.D., F.A.C.P.

Each month as authors update their contributions, revisions are entered on the magnetic tape on which the text and index are stored. The tape drives high-speed phototypesetting equipment so that subscribers receive about eight new chapters and a new index every four weeks; a bulletin highlights new developments.

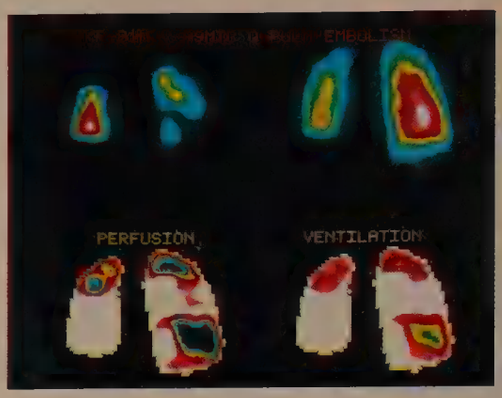
New material replaces old material in the living text, so that the information is there — up-to-date, at your fingertips.

A CME program of eight patient management problems offered over a 12-month period is available at no extra cost. As an organization accredited for continuing medical education, the Stanford University School of Medicine designates this continuing medical education activity as meeting the criteria for 32 credit hours in Category 1 for Educational Materials for the Physician's Recognition Award of the American Medical Association, provided it has been completed according to instructions. This program has been reviewed and is acceptable for 32 prescribed hours by the American Academy of Family Physicians. This program has been approved by the American College of Emergency Physicians for 32 hours of ACEP Category 1 credit.

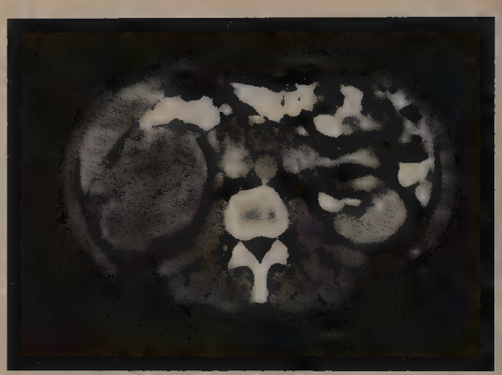
Trial Offer

We invite you to try SCIENTIFIC AMERICAN *Medicine* — for two months at no cost. Send us the coupon and you will receive the two-volume text and two monthly updates. You may also take a CME test for credit. At the end of 60 days, if you decide to continue the subscription, we will bill you for $220 for the full 12 months (renewal is currently $170); otherwise return the two volumes.

Please mail the coupon today and let us take the hassle out of keeping up.



Computerized scintigraphy reveals pulmonary thromboembolism.



Abdominal computed tomogram reveals large renal carcinoma replacing part of right kidney.

THE DISTINGUISHED AUTHORS AND THE FIFTEEN SECTIONS OF SCIENTIFIC AMERICAN ***MEDICINE***

1. Cardiovascular Medicine

Edgar Haber, M.D., Harvard Medical School and Massachusetts General Hospital

E. William Hancock, M.D., F.A.C.P., Stanford University School of Medicine

Roman W. DeSanctis, M.D., Harvard Medical School and Massachusetts General Hospital

Adolph M. Hutter, Jr., M.D., F.A.C.P., Harvard Medical School and Massachusetts General Hospital

Eve Elizabeth Slater, M.D., Harvard Medical School and Massachusetts General Hospital

2. Dermatology

Eugene M. Farber, M.D., Stanford University School of Medicine

Elizabeth A. Abel, M.D., Stanford University School of Medicine

3. Endocrinology

Daniel D. Federman, M.D., F.A.C.P., Harvard Medical School

4. Gastroenterology

Gary M. Gray, M.D., Stanford University School of Medicine

Peter B. Gregory, M.D., Stanford University School of Medicine

John Austin Collins, M.D., Stanford University School of Medicine

Douglas Wilmore, M.D., Harvard Medical School and Brigham and Women's Hospital

5. Hematology

Stanley L. Schrier, M.D., Stanford University School of Medicine

6. Immunology

John David, M.D., Harvard Medical School and Harvard School of Public Health

7. Infectious Disease

Thomas C. Merigan, M.D., Stanford University School of Medicine

Morton N. Swartz, M.D., F.A.C.P., Harvard Medical School and Massachusetts General Hospital

Cyrus C. Hopkins, M.D., Harvard Medical School and Massachusetts General Hospital

Adolf W. Karchmer, M.D., F.A.C.P., Harvard Medical School and Massachusetts General Hospital

Robert H. Rubin, M.D., F.A.C.P., Harvard Medical School and Massachusetts General Hospital

Harvey B. Simon, M.D., F.A.C.P., Harvard Medical School and Massachusetts General Hospital

Peter F. Weller, M.D., Harvard Medical School; Beth Israel Hospital, Boston; and Brigham and Women's Hospital

8. Intensive and Emergency Care

Edward Rubenstein, M.D., F.A.C.P., Stanford University School of Medicine

9. Metabolism

George F. Cahill, Jr., M.D., Harvard Medical School, Howard Hughes Medical Institute; and Brigham and Women's Hospital

10. Nephrology

Roy H. Maffly, M.D., Stanford University School of Medicine and Palo Alto Veterans Administration Medical Center

William Bennett, M.D., Health Sciences University, Portland, Oregon

11. Neurology

Robert W. P. Cutler, M.D., Stanford University School of Medicine

12. Oncology

Saul A. Rosenberg, M.D., F.A.C.P., Stanford University School of Medicine

13. Psychiatry

Ned H. Cassem, M.D., Harvard Medical School and Massachusetts General Hospital

14. Respiratory Medicine

Eugene D. Robin, M.D., F.A.C.P., Stanford University School of Medicine

15. Rheumatology

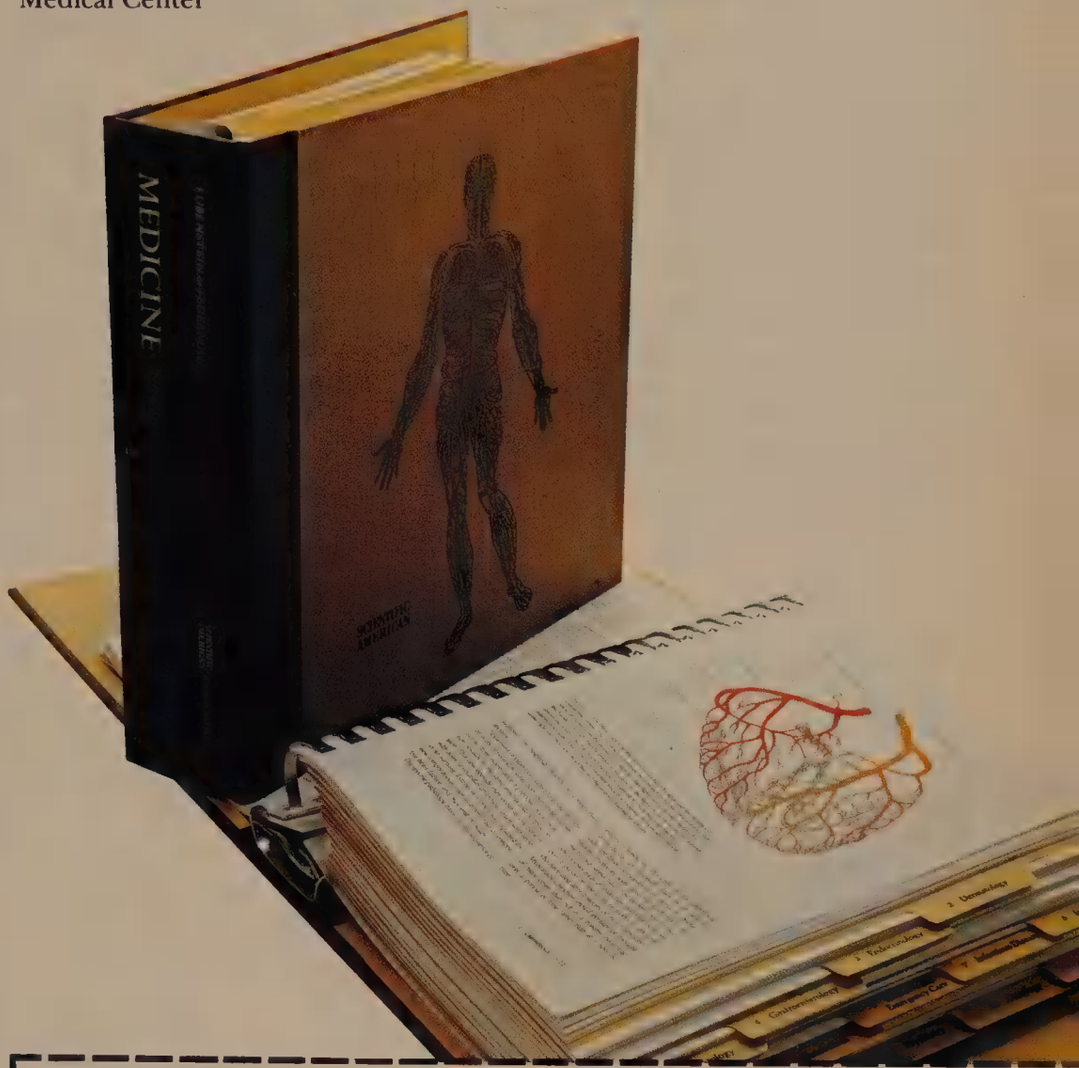
Stephen M. Krane, M.D., Harvard Medical School and Massachusetts General Hospital

Dwight R. Robinson, M.D., Harvard Medical School and Massachusetts General Hospital

Andrei Calin, M.D., M.A., M.R.C.P., Stanford University School of Medicine and Palo Alto Veterans Administration Medical Center

Order by Phone

You can order SCIENTIFIC AMERICAN *Medicine* by telephone. **Please call this toll-free number: 1-800-345-8112 (in Pennsylvania, call 1-800-662-2444);** you will be billed after your subscription begins. Toll-free calls are acceptable only for orders placed in the continental United States.



SCIENTIFIC AMERICAN MEDICINE

415 Madison Avenue, New York, N.Y. 10017

Please enroll me as a subscriber to SCIENTIFIC AMERICAN *Medicine*. On receipt of this coupon you will send me the advanced two-volume text described in your announcement and update it regularly by sending me new monthly subsections. I understand that the price of $220 for first year of service is tax deductible, as is current renewal at $170. If I am not entirely satisfied, I may cancel at any time during the first 60 days, returning all materials for a *complete refund.*

☐ Please enter my subscription for SCIENTIFIC AMERICAN *Medicine*
☐ I shall also enroll in the CME program
☐ I enclose a check made out to SCIENTIFIC AMERICAN *Medicine* for $220*
☐ Bill me ☐ VISA ☐ MasterCard

Expiration Date_____ Account Number_____

*Please add sales tax for California, Illinois, Massachusetts, Michigan and New York

Name_____

MD Specialty_____

Address_____

City_____ State_____ Zip_____

Signature_____

Please allow six to eight weeks for delivery. All payments must be in U.S. dollars. CME is available outside the U.S. and its possessions for an additional $65. Surface routes used unless airmail delivery is prepaid.

SCIENCE AND THE CITIZEN

Ill-starred Wars

President Reagan's televised appeal to "those who gave us nuclear weapons" to join in a new long-term effort to develop a nationwide anti-ballistic-missile (ABM) system has met with a largely negative response from the scientific community. The main arguments advanced in opposition to such an undertaking are much the same as they were in the 1960's, when the feasibility of ballistic-missile defense first emerged as a major public issue.

At that time the proposed method of dealing with a long-range missile attack on the U.S. by the U.S.S.R. called for the deployment of an extensive array of ground-based radars and nuclear-armed interceptor missiles. Critics of the idea pointed out that any system of this kind, no matter what the cost, was unlikely to be fully effective against a determined adversary and hence of dubious value given the great destructiveness of even a small fraction of the possible incoming nuclear warheads. The ABM system could not provide an impenetrable shield over the entire country because of various countermeasures available to the attacker. For example, the attacking country could change its offensive tactics, black out the system's long-range radars with high-altitude nuclear explosions, fool the defense with a large number of comparatively cheap decoys and other "penetration aids" or simply overwhelm it by adding enough offensive warheads to the attacking force to exhaust the supply of antimissile missiles. Such an action-reaction cycle in military preparations, it was argued, would lead inevitably to an intensification of the strategic-arms race [see "Anti-Ballistic-Missile Systems," by Richard L. Garwin and Hans A. Bethe; SCIENTIFIC AMERICAN, March, 1968].

The force of these arguments was formally recognized in the anti-ballistic-missile treaty of 1972, in which each superpower agreed "not to deploy ABM systems for a defense of the territory of its country and not to provide a base for such a defense." The treaty limited each side to local ABM systems at no more than two sites (later reduced to one).

Although President Reagan did not specify what kind of ABM system he had in mind when he spoke of his new "vision of the future," a spokesman for the Administration acknowledged that the systems currently under consideration include both particle-beam weapons and laser weapons, based either on the ground or in space. According to independent studies carried out at the Massachusetts Institute of Technology, these systems face insurmountable technological obstacles; moreover, they would be just as vulnerable to offensive countermeasures as the ABM systems of the 1960's were [see "Particle-Beam Weapons," by John Parmentola and Kosta Tsipis, SCIENTIFIC AMERICAN, April, 1979, and "Laser Weapons," by Kosta Tsipis, SCIENTIFIC AMERICAN, December, 1981].

The one thing that does seem certain about the president's initiative is how it will be perceived by military analysts on the other side of the nuclear "balance of terror." According to a statement signed by 244 scientists in the U.S.S.R., "there are no effective defensive means in nuclear war, and their creation is practically impossible." Nevertheless, the statement adds, the proposed "attempt to create so-called 'defensive weapons' to counter the strategic nuclear force of the other side... will inevitably lead to the emergence of a new element strengthening the American 'first strike' potential [by] striving to lessen the power of a retaliatory strike.... By his statement the President is creating a most dangerous illusion, which may cause an even more threatening spiral of the arms race. We are firmly convinced that this act will result in a sharp lessening of international security, including the security of the U.S."

Underground Science

Cosmic rays were once the principal source of particles for experiments in high-energy physics, but lately physicists have been going to considerable lengths—or considerable depths—to escape them. Experiments have been set up in mine shafts and tunnels, where they are shielded from most cosmic rays by the overburden of earth and rock. Indeed, in the past few years there has been a sudden proliferation of underground experiments, due in large part to interest in a search for the decay of the proton. The number of suitable deep sites, however, is quite limited. It has now been proposed to build a National Underground Science Facility, which would provide a permanent home and supporting staff for such work.

The need for cosmic peace and quiet is most acute in experiments that attempt to measure some exceedingly rare process whose signal might be lost in a noisy background of extraneous events. The conjectured decay of the proton is a notable example. The proton was long considered an immortal form of matter, but theoretical work in the past several years has given reason to believe a proton might occasionally decay. The average lifetime would have to be at least 10^{30} years, or some 10^{20} times the age of the universe, and a detector made up of thousands of tons of material could be expected to register no more than a few decays per year. They would have to be distinguished from the much more numerous events induced by the passage of cosmic-ray particles through the detector volume. Putting the apparatus deep underground minimizes this distracting and possibly misleading background.

Of the proton-decay detectors now in operation the largest is a tank filled with 10,000 tons of water installed 600 meters below the surface in a salt mine near Cleveland. It is operated by investigators from the University of California at Irvine, the University of Michigan and the Brookhaven National Laboratory. In January the group announced its first result: no decay events had been identified, and a new lower limit of 5×10^{31} years could be set on the average proton lifetime. For theorists the result suggests that certain simple and appealing models of the structure of matter may have to be excluded from consideration. For experimental physicists the result implies that a still larger and more sensitive detector may be needed to gather evidence of proton decay. Even if one of the existing experiments detects the disintegration of a proton, apparatus with finer spatial resolution will be needed to study the events in detail. An experiment of this kind would be a leading candidate for installation in a national underground science laboratory.

In the search for proton decay the most troublesome background events are those induced by neutrinos, particles that cannot be screened out by any feasible thickness of overburden; neutrinos readily pass through the entire earth. Several other proposed underground experiments would examine the neutrinos themselves. For example, the energy spectrum of cosmic-ray neutrinos is only poorly known and might be measured by an underground detector acting as a neutrino "telescope." The conjectured phenomenon called neutrino oscillation, in which one type of neutrino is converted spontaneously into another type, might also be detected.

Another neutrino experiment, proposed by Raymond Davis, Jr., of Brookhaven, would yield a refined measurement of the flux of neutrinos emitted by the sun. Among mine-shaft physicists Davis is a pioneer: for almost 15 years he has been counting solar neutrinos with a detector in the Homestake Mine in South Dakota. The active element in the detector is chlorine, which can be converted by a neutrino into argon. The results have been consistently puzzling (the measured flux of neutrinos is smaller than expected), and their interpreta-

Exxon introduces the incredible 965 Ink Jet Printer.

The automated office will never sound the same again.

From French to formulas. Dutch to double underline. Just about anything you can put in writing, Exxon's new 965 Ink Jet Printer can print ... quickly and quietly.

All thanks to the unique ink-jet technology designed by Exxon Office Systems to work with EXXON 500 Series Information Processors.

The new EXXON 965 Ink Jet Printer is pure breakthrough ... through and through. It's just about the quietest, most versatile printer you can buy.

It can print math and scientific symbols as easy as one, two, three. And even special oversized characters up to six lines high.

You can also print documents in eight on-line type styles of your choice, and all in italics ... simply by instructing the printer to do so.

The new EXXON 965 Ink Jet Printer can print just about anything at speeds of up to 90 characters per second. And all without disturbing a single person in the office.

If you'd like to know more about the incredible 965 Ink Jet Printer, just send in the coupon below. Or call **800-327-6666.**

After all, since you can't hear it, you'll have to see it to believe it.

EXXON OFFICE SYSTEMS

The future...without the shock.

SAM 0683

Exxon Office Systems
P.O. Box 10184, Stamford, CT 06904

I'd like to know more about

☐ the EXXON 965 Ink Jet Printer
☐ the EXXON 500 Series Information Processor

Please have your representative call.

Name _____ Title _____

Company _____

Address _____

City _____ State _____ Zip _____

Telephone _____

800-327-6666
CONNECTICUT 800-942-2525

tion might be clarified by a new experiment based on gallium.

Several more possible uses of an underground science facility have been discussed. The search for massive magnetic monopoles, for example, might benefit from a reduced background of cosmic rays. Attempts to detect gravitational radiation, or gravity waves, might exploit another characteristic of a deep underground site: freedom from vibration. In the stillness of the subterranean laboratory an instrument sensitive to low-frequency gravity waves (from one to 100 hertz) might be operated.

The chief proponents of the National Underground Science Facility have been Alfred K. Mann of the University of Pennsylvania and R. R. Sharp, Jr., and Leonard M. Simmons, Jr., of the Los Alamos National Laboratory. At a meeting held last September at Los Alamos, Mann reviewed the work that might be done in the facility and the technical issues that must be addressed in building it. The depth of the laboratory, he points out, should be sufficient to provide shielding from cosmic rays equivalent to at least 2,500 meters of water; because rock is denser than water, the actual depth could be appreciably less. Large rooms would be needed; one of them should be at least 50 by 200 by 50 feet. The main access shaft would have to be at least 12 feet in diameter.

Mann and Sharp have made a survey of the available underground sites in the western U.S. They identified 13 existing mines and tunnels that might be suitable for the laboratory. None of the sites is ideal, however, and some of the best ones are in working mines, where commercial operations would be likely to interfere with scientific activities. Attention has therefore focused on a plan to build an underground facility in a corner of the Nevada Test Site, the desert tract near Yucca Flat where the Department of Energy conducts underground tests of nuclear weapons. The area was set aside some years ago for nonclassified work. The proposed laboratory would be at a depth of 3,600 feet, equivalent to 2,800 meters of water.

Some physicists have objected to the idea of working at the Nevada Test Site. Nevertheless, Mann, Sharp and Simmons point out that the area has important advantages; indeed, it may have been chosen for weapons testing for some of the same reasons that make it suitable for an underground laboratory. The geology and hydrology of the site have been thoroughly explored and are exceptionally stable. The staff of the test site itself has expertise in deep excavation; more deep, large-diameter shafts have been drilled there, Mann says, than in any comparable area elsewhere in the U.S. Construction costs (estimated at $42.3 million) would be paid by the Department of Energy, but it is possible that not all the cost would come out of the department's basic-science budget.

A committee chaired by Norman F. Ramsey, Jr., of Harvard University has been formed to advise the director of the Los Alamos laboratory on matters related to science underground and the proposed facility. In a preliminary report the committee stated that a dedicated underground facility is urgently needed to provide space and technical support for important experiments in particle physics and astrophysics. The committee also expressed approval of the Los Alamos proposal to construct the facility at the Nevada Test Site.

Messenger with Forked Tongue

In the three decades since James D. Watson and Francis H. C. Crick elucidated the structure of DNA the picture of how the genetic apparatus works has become steadily more complex. In place of the original simple dictum that "DNA makes RNA and RNA makes protein" is an intricate mechanism in

which many intermediary operations have an importance almost equal to that of the fundamental genetic events. Throughout this period of intensive revision in molecular genetics one assumption has remained intact, namely the assumption that DNA and RNA molecules are linear polymers. Now even that presupposition has been contradicted. In *Proceedings of the National Academy of Sciences* John C. Wallace and Mary Edmonds of the University of Pittsburgh report finding an RNA molecule whose chain is split into two branches.

The finding of Wallace and Edmonds is a by-product of the considerable body of work that has been done on the processing of RNA in the living cell. The components of the RNA polymer are the nucleotide bases adenine, guanine, cytosine and uracil. Each nucleotide includes a five-carbon sugar unit whose carbons are generally designated by numbers. The monomeric units of the polymer are connected by a phosphate group interposed between the 3′ carbon of one sugar and the 5′ carbon of the adjacent sugar. The sequence of the nucleotides in the chain constitutes the information for constructing a protein in the process called translation. The RNA molecule is assembled on a DNA template in the cell nucleus in the process called transcription; the RNA is then transported to the cytoplasm, where the translation is done.

Intervening between transcription and translation are several significant operations. These processing operations transform the raw RNA transcript, known as heterogeneous nuclear RNA (hnRNA) in animal cells, into mature messenger RNA (mRNA), the form employed in translation. Among the processing functions are the addition of different structures to the ends of the hnRNA. A sequence of from 150 to 200 adenine nucleotides is attached to one end; the resulting molecule is designated poly(A)$^+$hnRNA. A structure called a cap, consisting of a modified guanine nucleotide, is attached to the other end of the molecule by a linkage that includes three phosphate units.

The attachment of the poly(A) sequence and the cap is thought to have a role in preparing the RNA for translation. The specific contribution made by each of these modifications, however, is not well understood. Wallace and Edmonds were investigating the detailed structure of the cap when they made their recent finding. The cap can readily be separated from the rest of the hnRNA by means of the enzymes called ribonucleases, which sever the covalent bonds linking adjacent nucleotides and thus yield the monomers. The chemical structure of the cap protects it from degradation by ribonucleases. Hence the application of the enzymes to hnRNA leaves an assortment of individual nucleotides and the intact cap.

Both the cap and the nucleotides have a negative electric charge. The charge of the cap, however, is much larger than that of the monomers. The two kinds of molecule can therefore be separated by ion-exchange chromatography, a technique that can be employed to separate molecules on the basis of differences in charge. Wallace and Edmonds applied the procedure to ribonuclease-treated hnRNA taken from the cancer cells designated HeLa cells. As expected, the procedure revealed the presence of the single nucleotides and the cap.

There was also a second molecular structure with a negative charge roughly equal to that of the cap. At first Wallace and Edmonds assumed the new substance was a kind of cap that had not previously been observed. Further chemical tests showed, however, that it was not a cap. Instead the sequence of tests demonstrated that the newly discovered segment of RNA includes a junction resembling the crotch of a tree

WHO BUT

ONLY BEEFEATER® GIN COULD MAKE ALL THESE STATEMENTS. AND BE SO ENJOYABLE.

"The Crown Jewel of England."

It actually has roots all over the world. Our juniper berries, for instance, come from the southern Alps, because that's where the choicest juniper grows.

Our angelica comes from Belgium. And our orris from Italy.

Are these trips necessary? Only if we (and you) insist on the world's most excellent gin.

The Master Stillman.

In London–home of the world's driest, most delicate gins–stillmanship is held as art, and our stillmen as masters of it.

Indeed, as you will see, there is in the distilling of Beefeater Gin a distinct resemblance to the art of the great jewelers.

We're all heart.

A distilling run consists of an unblemished middle, or "heart," and of "foreshots and feints," the beginning and end of the run, which are flawed and out of balance.

Our master distiller, like a master jewel-cutter, selects only the choicest flawless heart of the run–thus assuring you of a gin of delicate taste and impeccable balance.

Proof Perfect.

Beefeater Gin is bottled at precisely 94 proof.

BEEFEATER®?

is proof was not chosen by whim.
At an even slightly higher proof, the gin would
a shade less delicate. Any lower and it would be
race less flavorful. Our exacting proof imparts to
efeater a balance between delicacy and flavor
ely achieved in any other gin.

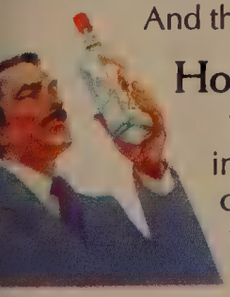
The Legacy of James Burrough.

Our founder had a basic philosophy: if you want
hing done right, do it yourself.

Not a single distillation of his
efeater Gin left the distillery until
was approved by Mr. Burrough
mself.
He wasn't much of a delegator,
t he made an exceptional gin.
His philosophy has been
nded down through succeeding
nerations of the Burrough family, who still approve
ch batch of Beefeater before it can leave the
stillery.
The apple never falls far from the tree. For which
drinkers to this day are profoundly grateful.

London is as London does.

A lot of gins have the word "London" on their
bels. But that does *not* necessarily mean they
re distilled in London.
In fact, there is only one major gin you can buy
the United States that is distilled, bottled and
aled in London.
And that gin is Beefeater Gin.



Hold Beefeater up to a light.

That's what we do at the distillery in London. The bottle is absolutely clear, rather than colored glass, so you can appreciate the brilliant clarity of Beefeater.

We've got your number.

Look, for example, at the individual registration
mber on the back label of any bottle of Beefeater
n. If it happened to be the
e shown here, it would
ll you that your Beefeater
as distilled Thursday, January 7, 1982, under the
pervision of Thomas Bowman, Master Stillman,

No. JF 1222146

and that this particular distillation was approved personally by Mr. Norman Burrough, a direct descendant of our founder, before it was allowed to leave the distillery.

Does it do any good to know all this? It does indeed. That number is a vital part of Beefeater Gin's quality control.

Because the number is Mr. Bowman's personal signature. And Mr. Burrough's as well.

And any artist is at his best when he has to sign his work.

A Most Exclusive Honor.

Enjoyed in over 170 countries, Beefeater Gin has achieved such recognition around the world that its makers have been honored four times with The Queen's Award…presented on the birthday of Her Majesty, Elizabeth II. No other gin has earned this coveted award even once.

First Name for the Martini.

It is in the Martini that the differences among gins become most evident.

There is nothing in a Martini, you know, except the gin itself, a dash of vermouth, and perhaps an olive.

For a gin, then, the Martini is the moment of truth. Because the gin in a Martini stands virtually alone for you to judge.

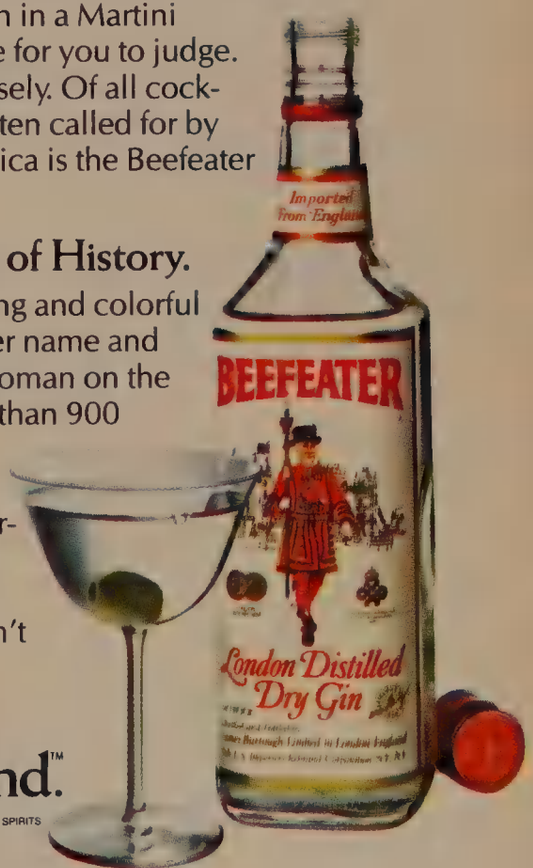
And you judge wisely. Of all cocktails, the one most often called for by brand name in America is the Beefeater Martini.

The Lessons of History.

Beefeater has a long and colorful history! The Beefeater name and the scarlet-coated yeoman on the bottle go back more than 900 years, to the regiment that guarded William the Conqueror in the Tower of London.

Of course, you can't really *taste* history.

Or can you?



BEEFEATER® GIN. The Crown Jewel of England.™

IMPORTED FROM ENGLAND BY KOBRAND CORP., N.Y., N.Y. 94 PROOF 100% GRAIN NEUTRAL SPIRITS

with two branches extending from it. The central member of the junction is a single nucleotide. As in linear RNA, strings of nucleotides extend from the 3′ and 5′ carbons of the sugar unit. In addition an anomalous third string of nucleotides extends from the 2′ carbon.

Wallace and Edmonds note that this is the first report of a branched RNA molecule. Since branching is found only in the hnRNA and not in the mature mRNA, the branches could play a role in RNA processing. Indeed, Wallace and Edmonds go on to speculate that the branching could contribute to one of the major steps in processing. Unprocessed hnRNA includes two kinds of base sequence. One kind corresponds to the parts of the gene called exons, which are ultimately translated into protein. The other kind corresponds to the parts of the gene called introns, which are not translated. In the assembly of the finished mRNA the introns are removed and the ends of adjacent exons are spliced. Consider a piece of hnRNA with two exons separated by an intron. The authors note that if the RNA were cut at one end of the intron, the free end of the adjacent exon could be brought to the other end of the intron, where it joins the second exon. The free end could then be attached to the 2′ carbon of the first nucleotide of the second exon. Such an operation would yield the branched structure observed by Wallace and Edmonds. Later the dangling intron attached to the 3′ carbon could be removed and the spliced exon could be transferred from the 2′ carbon to the usual 3′ site.

Clouded Ball

Cold clouds of neutral hydrogen atoms have a special place in the conduct of modern astronomy. Detected with radio telescopes by means of their radiation at a wavelength of 21 centimeters, they have proved to compose as much as a fifth of the mass in a typical spiral galaxy. Thus they are excellent markers of a galaxy's spiral structure. In recent months Yervant Terzian, Stephen Schneider, George Helov and Edwin Salpeter of Cornell University have employed the radio telescope at Arecibo in Puerto Rico to detect clouds of neutral hydrogen atoms in galaxies other than our own galaxy. In January they aimed the telescope toward a point between two small groups of galaxies in the constellation Leo. They intended only to calibrate their instruments by measuring the "background" at 21 centimeters when the telescope was pointed at "blank sky." Instead they found an immense cloud of neutral hydrogen in intergalactic space.

The intergalactic cloud is 10 megaparsecs from the earth (one megaparsec is 3.26 million light-years). It is at least 100 kiloparsecs across, or about three times the diameter of our galaxy. Its content of hydrogen atoms is at least 10^9 times the mass of the sun. Optical images of the sky show no sign of stars between the two groups of galaxies, but then the density of the hydrogen in the intergalactic cloud is only from 10^{-3} to 10^{-4} atoms per cubic centimeter. The spontaneous formation of stars by the gravitational collapse of the matter in a cloud would require a density of 10^4 or even 10^6 atoms per cubic centimeter.

On the other hand, the western edge of the cloud is receding from the solar system (as part of the expansion of the universe) faster than the eastern edge. The difference in velocity, which is about 80 kilometers per second, can be taken as evidence that the cloud is rotating as it recedes, and the rotation in turn implies that the cloud is kept from flying apart by the gravitational effect of undetected matter amounting to as much as 10^{11} solar masses, or 100 times the mass of the detected hydrogen atoms. The cloud would then be equal in mass to a galaxy.

Picolife

Exceedingly small plant life is proving to be surprisingly important in the ecology of the oceans. At two sites in the tropical Pacific a group from the Marine Ecology Laboratory of the Bedford Institute of Oceanography in Nova Scotia sampled the ocean water each day for six days and passed the samples through various filters. The intent was to trap picoplankton: organisms whose wet weight is on the order of a picogram, or 10^{-12} gram. W. K. W. Li and his colleagues defined picoplankton as cells that pass through a filter whose pores are one micrometer in diameter but do not pass through a filter with .2-micrometer pores.

The cells that met the criteria were found to contain as much as 90 percent of the chlorophyll in the water samples. Thus many such cells are plants. It had earlier been supposed that picoplankton must be predominantly heterotrophic bacteria (bacteria whose mode of nutrition is not photosynthetic). A technique in which chlorophyll is made to fluoresce red or orange under the microscope revealed as many as 1.5 million picoplanktonic plant cells in each milliliter of seawater in the uppermost 50 meters of the oceanic water column.

Evidence that the picoplanktonic plants contribute to the photosynthetic fixation of carbon at the base of the ocean's food chain came from further experiments, which Li and his colleagues describe in *Science*. In one experiment "more than 70 percent of the total measurable activity of the enzyme ribulose biphosphate carboxylase was found in the picoplankton." The enzyme is present in organisms that derive organic carbon photosynthetically from inorganic sources such as carbon dioxide. In another experiment cells that had passed through the one-micrometer filter were given inorganic carbon labeled with the radioactive isotope carbon 14. The rate at which radioactive carbon accumulated in the cells turned out to be substantially less in the dark than it was in the light.

The experiments "support a view of the structure of the pelagic ecosystem in which most of the activity (production and metabolism) is carried on by very small organisms." It is thought that the oceans are inhabited by forms of life even smaller than picoplankton, weighing about as much as a virus. Essentially nothing is known about them.

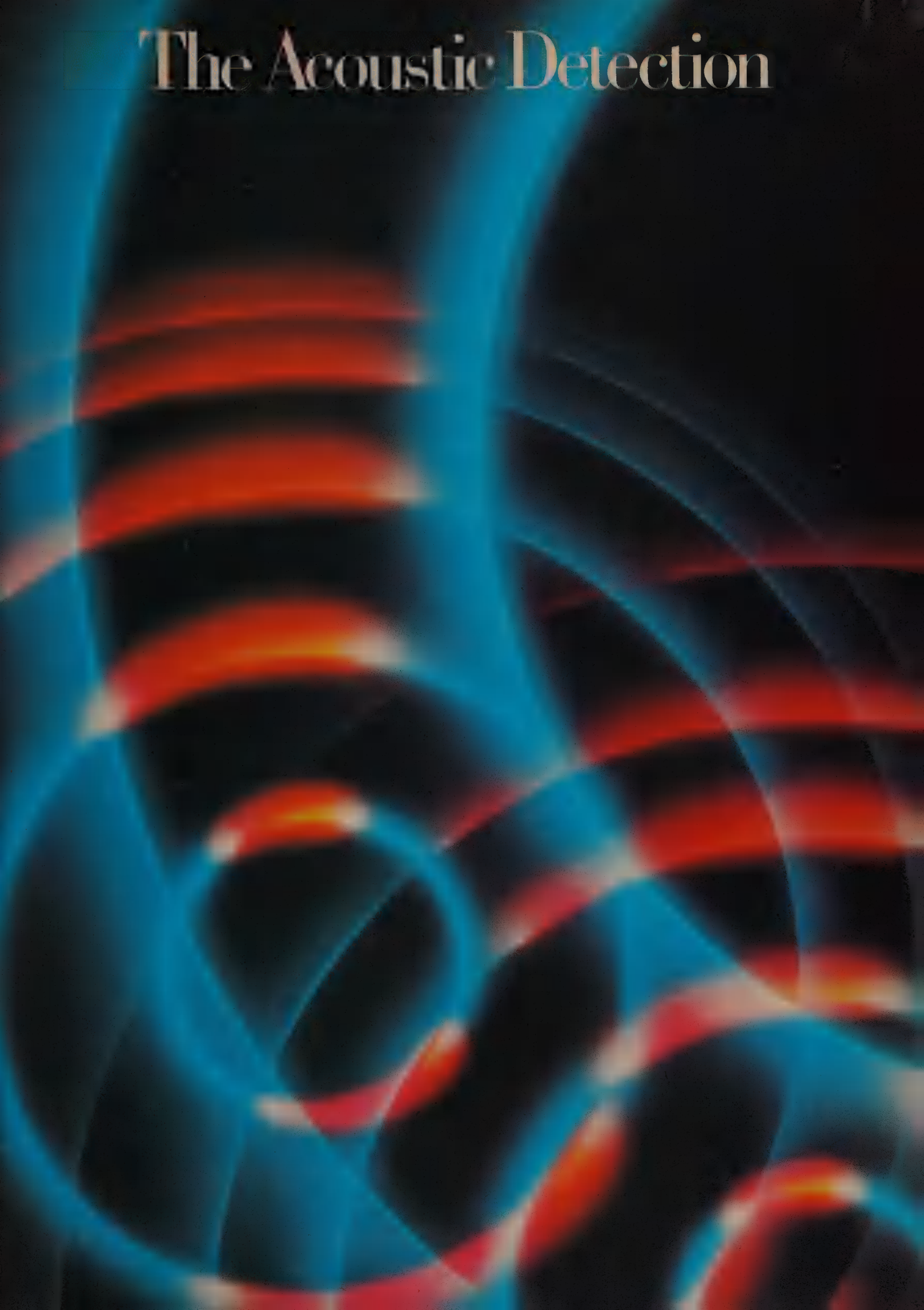
Spinning Still

A standard industrial distillation column operates with no moving parts to separate two or more components of a liquid on the basis of their different boiling temperatures. The process is well understood and reliable, but the capital cost of the column and its associated equipment is high. Engineers at Imperial Chemical Industries in the U.K. are proposing a new distillation process in which the separating unit is a spinning drum. They think the technique may offer several advantages over the standard system, including lower capital costs.

In the conventional process for distilling a substance such as petroleum a stream of heated fluid is pumped into the lower part of a fractionating column as a mixture of liquid and vapor. The column, which may be as much as 150 feet high, contains perhaps 40 fractionating trays spaced regularly through its height. The vapors rise through the column and at the top are condensed by cooling. Part of the condensed liquid (representing the component with the lowest boiling temperature) is taken off. The rest goes back down through the fractionating trays, coming in contact with rising vapor at each level. At each stage the liquid absorbs heavier constituents from the vapor and loses lighter ones. When a steady state has been reached, the column is separating all the constituents, the ones with a lower boiling point at the top and the ones with a higher boiling point at the bottom.

In the new Higee (for high gravity) system the key unit is a flat, round drum filled with the kind of porous packing material installed in many distillation towers. The drum is spun at about 1,800 revolutions per minute, generating a centrifugal force of about 900 *g* (900 times the gravitational force at the surface of the earth). The heated liquid is pumped into the center of the drum and flows outward by centrifugal force. A countercurrent of vapor enters through holes in the rim and flows inward owing

The Acoustic Detection



The Acoustic Detection

Scientists have studied microstructural discontinuities in high-carbon steel since the early 1920s. By monitoring acoustic emissions, a materials research engineer at the General Motors Research Laboratories has arrived at a more detailed understanding of how one type of discontinuity occurs.

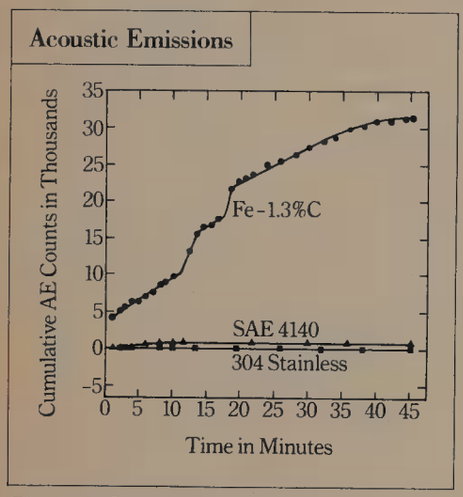
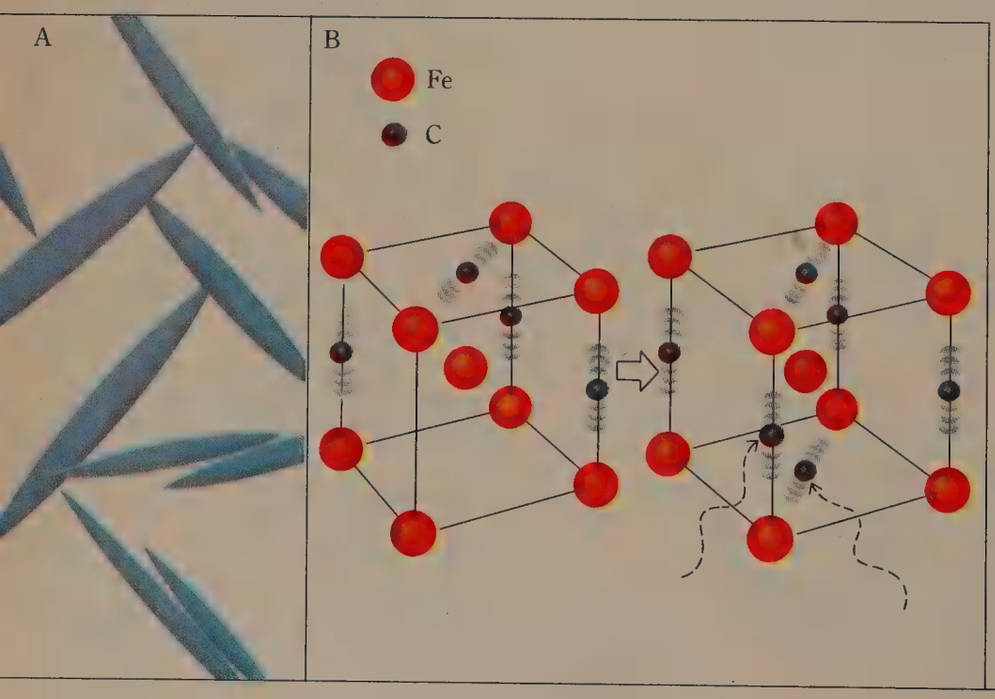


Figure 1: Cumulative acoustic emission counts for Fe-1.3%C steel, and control specimens of SAE 4140 and 304 stainless steel.

Figure 2: Artist's rendering of two proposed sources of microcracking: (A) impingement of the plates during the formation of martensite and (B) carbon atom rearrangement during the aging of martensite.



MARTENSITE is a hard microconstituent of steel which forms when austenite, iron containing carbon in solid solution, is quenched from a high temperature. The martensitic transformation produces steel that is hard and strong, but non-ductile. Through heat treatment, the steel can be tailored to applications requiring different degrees of ductility. High-carbon martensite—a highly stressed microstructure with a plate-like morphology—contains microscopic ruptures or separations 10 to 20 microns in length. These structural discontinuities, termed "microcracks," influence the mechanical properties of steel.

Although aspects of the microcracking phenomenon have been understood by metallurgists for more than fifty years, there is still no definitive explanation for when or how it occurs. An engineer at the General Motors Research Laboratories has devised an experiment that detects the microcracks as they occur.

The elastic energy released when microcracks form should produce a stress wave and associated high-frequency acoustic emission (AE). Using a piezoelectric transducer as the monitoring device, Dr. Michael Shea set out to determine what could be learned about the microcracking process by measuring AE.

The more widely accepted of two current hypotheses—the "impingement model"—asserts that microcracking is transformation-induced, taking place due to the collision of martensite plates during the quench. The other model maintains that microcracking occurs during the aging of martensite after the plates have already formed. The "aging model" suggests that thermal activation enables carbon atoms to rearrange themselves, producing localized stresses high enough to cause microcracking. Dr. Shea's ongoing research into high-carbon martensite led him to believe that the aging hypothesis was important. He proceeded to determine if AE is produced during aging.

For his study, Dr. Shea chose Fe-1.3%C steel, which undergoes martensitic transformation during quenching and is known to form

microcracks. To provide baseline data, control specimens of 304 stainless steel and SAE 4140 steel were put through the same procedures as the test composition. When quenched, 304 stainless steel produces no martensite, and SAE 4140 forms a low-carbon martensite which has a lath-type morphology, and generally does not microcrack.

SPECIMENS of the three compositions were quenched to -196°C and then slowly heated to room temperature. Acoustic measurements were made beginning at 0°C, at which point carbon atom mobility is sufficient to allow rearrangement processes to take place, and continued for 45 minutes after the specimens had reached room temperature. No AE was recorded for 304 stainless steel, and only a slight amount for SAE 4140. Significant emission, however, was measured for the Fe-1.3%C steel specimen during the entire testing period (see Figure 1). Since martensite had already formed during the quench, these results support the hypothesis that microcracking is produced during aging of the freshly-formed plates. Dr. Shea ruled out both slip and twinning as sources of AE since the literature indicates that neither factor is significant during aging of martensite below 40°C. The possibility that the AE resulted from isothermal transformation of austenite to martensite could also be excluded because this process does not take place in the composition studied.

"These results demonstrate conclusively," says Dr. Shea, "that microcracking occurs during the aging of high-carbon martensite, thereby providing support for the less accepted of the two models.

"The next challenge," he continues, "will be to quantify the relative contributions of both models—impingement and aging—in an effort to determine which, in fact, is the more important mechanism, thus furthering our understanding of microcrack formation. Then, perhaps, we can more systematically explore ways to minimize microcracking."

General Motors

GM

MARK OF EXCELLENCE

THE MAN BEHIND THE WORK

Dr. Michael Shea is a Staff Research Engineer in the Metallurgy Department at the General Motors Research Laboratories.

Dr. Shea received his undergraduate and graduate degrees in metallurgical engineering from Michigan Technological University, and his Ph.D. in materials engineering from Rensselaer Polytechnic Institute. His thesis concerned deformation and fracture of cesium chloride type superlattices. He joined General Motors in 1971.

The areas of metallurgical research pursued by Dr. Shea at General Motors include the mechanical properties of high-carbon steels, mechanically-induced transformation of austenite, and structure/property relationships in nodular cast iron. His exploration of the microcracking phenomenon in martensite was conducted with the help of instrumentation developed by GM colleague Dr. Douglas Harvey.

to a higher pressure at the outside. In the porous packing the liquid and the vapor flow past each other in thin layers, making for a rapid separation.

A demonstration unit set up by Imperial Chemical Industries in 1981 operates with two rotors, each rotor about 30 inches in diameter and six inches thick. It has successfully separated ethanol and isopropanol over extended periods of time. These liquids boil respectively at 78 and 82 degrees Celsius and would require a conventional distillation column about 100 feet high and three feet in diameter for a separation of comparable quality. The developers say their system can be cleaned, repaired or moved easily. What remains to be seen if it is scaled up to industrial size is whether the maintenance costs can be kept within reasonable bounds.

Sharing the Schlepping

Why do car pools tend to fall apart? Four co-workers agree that in principle it makes sense to share the driving, the fuel and the space on the highway for the trip to work and back. On Monday, however, Jones cannot take her turn because she must be out of town. On Tuesday, Smith must drive to work because he needs his car at lunchtime, even though his turn does not come until Thursday. Jones trades with Brown, who is scheduled to drive on Wednesday, but Roberts, who should drive on Tuesday, cannot trade with Smith and drive on Thursday. Jones is delayed returning from her trip, and so she cannot drive on Wednesday either. Smith refuses to drive three days in a row, so that either Brown or Roberts must drive on Wednesday. By the following week no one can remember who is to drive when.

Ronald Fagin and John H. Williams of the San Jose Research Laboratory of the International Business Machines Corporation have now devised an ingenious system for determining who is to drive on any given day. The system allows the members of the car pool to miss a turn, to drive out of turn or to trade driving days with one another at will. More important, it sets up a simple bookkeeping method that fairly accounts for the contribution each person makes to the car pool. Since the system is fair and flexible and leaves no doubt about whose turn is next, it can be applied whenever the people in a group agree to take turns doing some regular task, which might well be cooking rather than driving.

Fagin and Williams describe their carpooling algorithm in the *IBM Journal of Research and Development.* Suppose the maximum number of people who can ride together in one car is N, and U is the least common multiple of all the numbers less than or equal to N. Thus if there are three people in the car pool, U is equal to six; if there are four people, U is 12. The bookkeeping system worked out by Fagin and Williams maintains a running score for each member of the car pool. Points are added to a member's score for driving and subtracted for riding; the score remains unchanged when a member misses a day or drives alone. If the number of participants on a given day is K (which can be less than N), the driver on that day is given $U(K-1)/K$ points, and U/K points are deducted from the running total of each rider. Hence if three out of four people ride together, the driver receives $12 \times 2/3$, or eight, points and the two riders each lose $12/3$, or four, points. Each day the participant with the fewest points is selected as the driver.

Fagin and Williams point out that simpler token systems, such as the system in which a driver always receives one ride token from each rider, are inherently unfair. A person whose turn to drive came up repeatedly on sparsely attended days would receive less credit than someone who drove when the car was full. A better system would require, say, Jones to drive a third of the time she shares a ride with two other people and half of the time she rides with only one other person. The new system is fair in an even more flexible sense. The total number of times Jones must drive is equal to the sum of half of her driving share for two-person car pools, a third of her share for three-person car pools and a fourth of her share for four-person car pools. Thus she need not drive half of the time she shares a ride with one other person because the system shows how she can make up the deficit in the larger car pools. Moreover, the car pool need not be confined to one car; each person's contribution to car pooling could be fairly credited, say, throughout an institution.

The Big Picture

It was air travelers in the 1940's who called attention to the remarkable large drawings on the *pampa colorada* of Peru, a gravel-covered region of coastal red desert near Nazca some 300 kilometers south of Lima. Ever since, the who, when and how of the drawings' origin have fired the imagination of various commentators and tested the ingenuity of soberer scholars. Thanks largely to the patient field work of Maria Reiche, a German-born mathematician who lives in Peru, some of the wilder proposals have been laid to rest. These ranged from attributing the drawings to the work of extraterrestrials marking their desert spaceports to the claim by archaeoastronomers that the drawings provided sight lines for the observation of celestial phenomena.

Archaeological studies of the *pampa colorada* have identified both the who and the when of the Nazca figures. Pottery fragments remaining in the ancient work camps of the people who made the drawings and the carbon-isotope age of charcoal from the same workers' hearths unequivocally indicate that the vast desert displays were made by Nazca Indians sometime before A.D. 600.

The question of how is easily answered in part: the drawings were formed by the painstaking removal of gravel from the desert surface, exposing the lighter-colored underlying soil. The troublesome point has been how the drawings were laid out; many of them are too large to be seen in their entirety from ground level. Even the knowledgeable Reiche has suggested that the Nazca must have possessed surveying "instruments and equipment... buried and hidden from the eyes of the [conquistadors] as the one treasure that was not to be surrendered."

Writing in *The Skeptical Inquirer,* Joe Nickell of the University of Kentucky reports his doubt that the Nazca needed much surveying skill or special instrumentation to construct the desert drawings. He goes on to recount his success in replicating a 440-foot-long portrait of a condor using only limited and simple equipment. Nickell and his colleagues prepared a scale drawing of the Nazca original (Reiche had long ago found that the Nazca figure-makers had made such small mock-ups for guidance), determined a center line and plotted the coordinates of various points on the figure. Any given number of units on the small drawing was then translated into the same number of larger units (about 12.7 inches long) on the full-scale figure. Taking advantage of a fresh landfill area in West Liberty, Ky., the survey group used a rope subdivided into 100 Nazca "feet" to locate various points on the condor figure. The only surveying instrument they required was a wood cross to ensure that measurements were made at a right angle to the center line.

The survey group staked out 165 points and connected them with twine to duplicate the main outlines of the condor. Where a circle was required (for example in the representation of the condor's head) it was scribed with a stake and rope; all other curves were done freehand. After completing the outline the group marked it by hand on the dark soil of the landfill with powdered lime and arranged to have the replica photographed from an altitude of 1,000 feet. The match with the Nazca original was remarkable in its exactness. The ground drawing may possibly have been the world's largest art reproduction (until rain washed away the lime). Summarizing his group's experience, Nickell emphasizes that the task required no materials not readily available to the Nazca.

WHY CONTROLLERS OF BIG CORPORATIONS AND PRESIDENTS OF SMALL COMPANIES STAY AT HYATT.

People in business who understand value have something in common. They appreciate all the touches of Hyatt.®

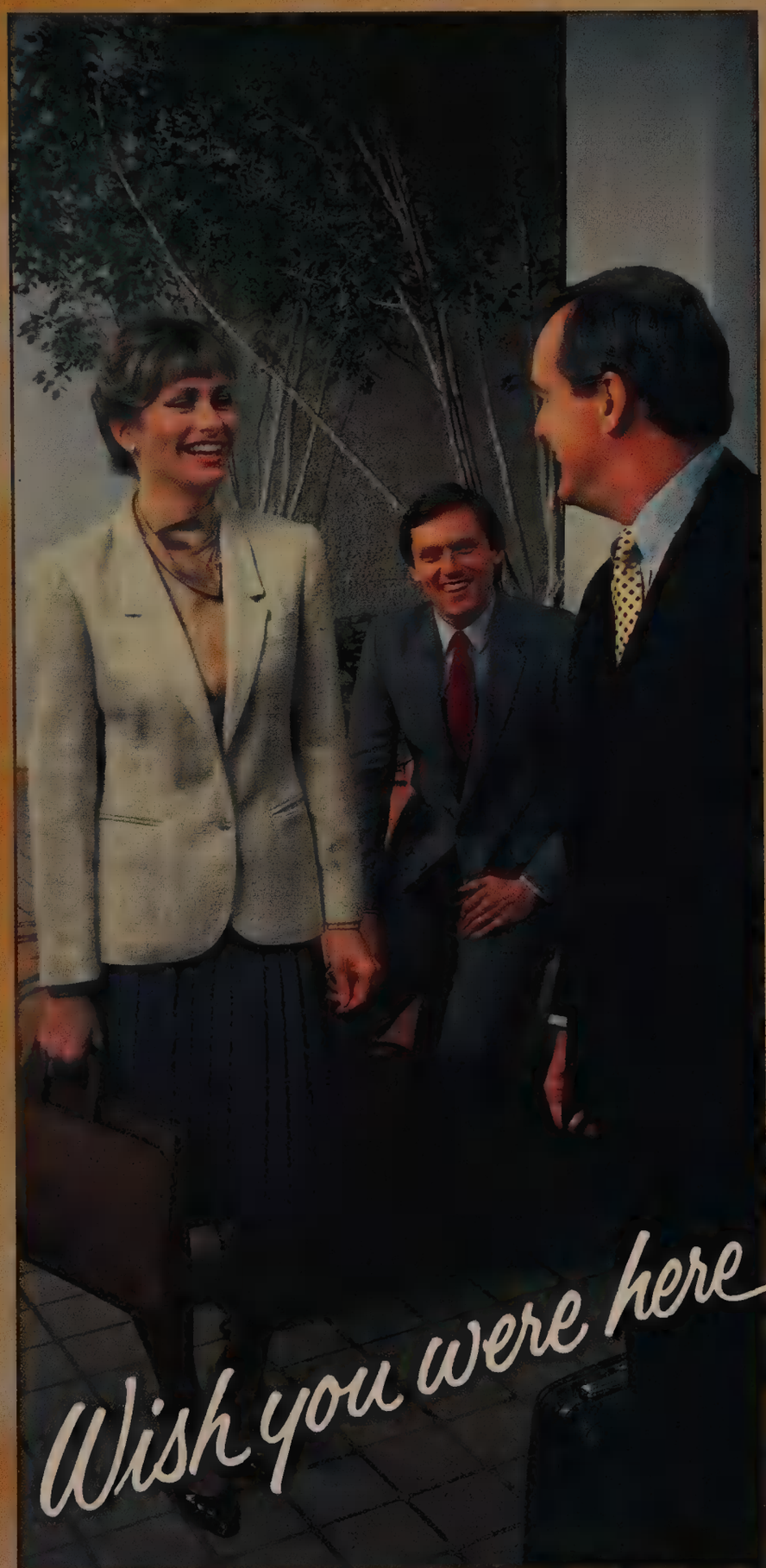
The soaps and the mints. The free morning newspaper. The quick, easy check-in. And the express check-out which allows them to be billed later instead of being kept waiting in line.

But most of all they appreciate the fact that they could be staying somewhere else, and getting so much less.

Spending money on a hotel that doesn't make you feel your best is an expense no business can afford. Controllers of big corporations and presidents of small companies understand that.

That's why they stay at Hyatt.

For reservations at 109 hotels worldwide, call your travel planner or 800 228 9000.



ATLANTA Discover magnificent Peachtree Center, 15 minutes from Atlanta International Airport.

CHICAGO O'HARE Hyatt has the city's vitality with the advantages of airport location.

DALLAS Soar above the lively downtown Reunion area at Hyatt.

MINNEAPOLIS Hyatt is downtown on Nicollet Mall just 2 blocks from corporate business district.

NEW YORK Grand Hyatt on fashionable Park Avenue at Grand Central.

©1983 Hyatt Hotels Corp.

HYATT® HOTELS

Peking Man

In 230,000 years of cave dwelling Homo erectus pekinensis left an abundance of fossils and artifacts that now give information on biological, technological and perhaps even social evolution

by Wu Rukang and Lin Shenglong

One of the sites with the longest history of habitation by man or his ancestors is a cave near the railroad station in Zhoukoudian, a town some 50 kilometers southwest of Beijing. Whereas modern cities are generally no more than a few thousand years old, the cave at Zhoukoudian was occupied almost continuously for more than 200,000 years. The multiple layers of fossil-bearing deposits indicate that early men first took shelter in the cave 460,000 years ago, and the last of them did not abandon the site until 230,000 years ago, when they were forced out by the filling of the cave with rubble and sediment. The species of mankind that lived there is classified as *Homo erectus pekinensis,* or Peking man. (The name became established before the customary English spelling of the city was changed to Beijing.)

The long record of habitation at Zhoukoudian offers an opportunity to trace the development of a single community over a period that spans a significant fraction of the evolution of the genus *Homo.* The period is long enough for progressive changes in the form of the fossils themselves to be discerned; one of the important changes in physical features is an increase in cranial capacity. Equally important, it is possible to reconstruct certain events in the cultural evolution of the species. There is evidence in the cave deposits that Peking man was able to control fire and that he employed it for cooking. The fossilized remains of animal bones indicate that the Zhoukoudian cave dwellers were effective hunters of both large and small game; fossilized seeds suggest another component of the diet. An abundance of flaked-stone implements provides information on toolmaking skills. From an analysis of the materials recovered from the cave it is even possible to speculate on the social organization of the community. For example, there is information to support conjectures about the sharing of food and the division of labor between the sexes.

In the past five years a comprehensive investigation of the Zhoukoudian site has been undertaken by more than 120 Chinese scientists, including us. The work is sponsored by the Institute of Vertebrate Paleontology and Paleoanthropology of the Chinese Academy of Sciences, and the contributors represent 17 universities and research institutions. Their fields of expertise range from early man and his artifacts to the study of ancient climate, caves and the terrains in which they form, ancient soils, pollen and the dating of archaeological materials. Here we shall present the salient findings of the group and attempt to sketch a portrait of Peking man in his native environment.

In the overall scheme of higher primate evolution Peking man and the other examples of *Homo erectus* are a comparatively late development. The first hominoid, the ancestor of both man and the anthropoid apes, had branched off from the other primates by about 35 million years ago. It walked quadrupedally and relied on seeds and fruits for its food. From 10 to eight million years ago the first hominid, the founder of a genetic lineage whose only living representative is modern man, diverged from the other hominoids such as the chimpanzee and the gorilla. The most primitive hominids apparently had a small cranium, with a capacity of perhaps 350 cubic centimeters; they could walk on their hind limbs and may have been able to carry objects while walking. It was not until about four million years ago that an advanced hominid, classified in the genus *Australopithecus,* appeared. The earliest remains of australopithecines have been found in the middle Awash River valley of Ethiopia. The skeleton was evidently capable of supporting a fully bipedal gait and the brain case had a capacity close to 500 cubic centimeters.

In the past 100 years many fossils of manlike skulls and skeletal fragments more advanced than those of *Australopithecus* have been found in Europe, Asia and Africa. For a time almost every discovery gave rise to a new taxonomic division, but virtually all these later specimens have now been subsumed in the single species *Homo erectus.* On the basis of chronological studies of the geologic context in which the fossils were found it has been ascertained that the species emerged at least 1.5 million years ago. *Homo erectus* could walk upright, as the species name suggests, but his skull was still very primitive, with a cranial capacity ranging from 850 to 1,100 cubic centimeters. He was able to make more advanced stone implements than the pebble tools of earlier hominids. Some later specimens of *Homo erectus,* including Peking man, developed a quite elaborate culture characterized by the inhabiting of caves and the hunting of game. They were also able to make use of fire to cook food, although it is not yet clear whether they were actually able to kindle a fire.

The subsequent course of development from *Homo erectus* to modern man cannot be traced in detail. Suffice it to say that about 200,000 years ago there appeared a form of man with a less heavily built face and a larger brain case than those of *Homo erectus.* The members of this species, early representatives of *Homo sapiens,* came to flourish in Europe, Asia and Africa somewhat less than 100,000 years ago. They were capable of building shelters in the open air and of starting a fire, and they made highly refined stone tools. Because of the closeness of their morphology to that of modern man and the progressiveness of their culture, they are considered to be among the direct ancestors of the modern *Homo sapiens.*

The Cave of Zhoukoudian

The cave in Zhoukoudian where the remains of Peking man have been found is in a limestone formation called Dragon-Bone Hill. Before it was known as a site of early man it had long been a favorite haunt of dragon-bone collectors. Dragons had an important place in traditional Chinese culture, and their bones (actually the fossils of various mammals) were thought to have great

medicinal value. Indeed, it was the dragon-bone collectors who found the first evidence of Peking man, although they did not recognize its significance.

At the beginning of the 20th century a number of paleontologists and anthropologists visited Zhoukoudian in search of evidence of primitive man. Notable among them were J. G. Anderson, a Swedish geologist, and Walter W. Granger, a paleontologist of the American Museum of Natural History. On a routine survey of Dragon-Bone Hill in 1921 Anderson, Granger and O. Zdansky, an Austrian paleontologist working as Anderson's assistant, were directed by local residents to a cave that was said to be "full of dragon bones." Nothing of interest was found, however, except some quartz fragments; because the fragments were far from their geologic context, Anderson considered them artifacts of early man.

Although survey excavations were undertaken in the cave, the progress of discovery was slow. In 1923 Zdansky found two anthropomorphic teeth. When a systematic excavation was started in 1927, Birger Böhlin, a Swedish paleontologist, discovered a well-preserved lower molar. Davidson Black, who was then professor of anatomy at Peking Union Medical College, assigned the fossil to the new species *Sinanthropus pekinensis.* From 1927 to 1937, when work was stopped by the Japanese invasion of China, the cave was continuously excavated. Pei Wenzhong, who was in charge of the Zhoukoudian excavation from 1928 through 1935, found the first complete cranium in the winter of 1929, thereby providing a sound basis for scientific studies of Peking man.

Since the founding of the People's Republic of China in 1949 the excavation of the cave has been resumed by the Institute of Vertebrate Paleontology and Paleoanthropology. Up to 1966 fossils of more than 40 males and females of various ages had been found associated with tens of thousands of stone artifacts and evidence of fire use. In addition the fossils of two species of plants (and the pollen of many more species) and 96 kinds of mammals were found. The abundance of the fossils and the thoroughness with which they have been studied make it possible to reconstruct elements of Peking man's 200,000-year history.

The cave, called formally Zhoukoudian Locality I, is a huge karst cave in a limestone structure formed in the Ordovician period, about half a billion years ago. Karst landscape, which takes its name from that of a region along the Dalmatian coast of Yugoslavia, is formed by the work of underground water on massive soluble limestones. The karst cave inhabited by Peking man is 140 meters long from east to west. The

LIMESTONE CAVE was inhabited by Peking man, a variety of *Homo erectus,* beginning 460,000 years ago and continuing until 230,000 years ago. The cave was discovered in the early 1920's. Excavations over the past 50 years and more have revealed evidence of the use of fire, an assemblage of 100,000 stone artifacts, fossilized bones and teeth of more than 40 individuals and fossil remains of animal and plant species that were presumably part of Peking man's diet. The photograph shows the east entry to the cave, which was the primary entrance in the first phase of Peking man's occupation. The entry was abandoned after a massive collapse of some 300,000 years ago.

widest part, which measures 40 meters, is at the eastern end; the narrowest part, at the western end, is only a little more than two meters wide.

The original cave, however, was not like this, nor was it habitable at the outset for Peking man. The formation of the cavern inside the hill began in the Pliocene epoch, about five million years ago, when water percolating along both horizontal and vertical cracks dissolved the limestone and carried it away in solution. The horizontal erosion created a long cave with both ends narrower than the central part; the vertical erosion followed precipitous fissures, creating funnel-shaped gullies whose bottom was deep, rough and bumpy. The original cave was a hollow in Dragon-Bone Hill without openings to the surface. Owing to the continual scouring of the hill by the ancient Zhoukou River, however, a spur on the eastern slope of the hill was cut off early in the Pleistocene epoch, exposing the eastern wall of the cave to water erosion. As a result the eastern slope developed a diagonal fissure approximately at the present-day location of the small east entry to the cave. The fissure became progressively wider and in the end established a connection between the Zhoukou River and the cave. Sand and dirt carried by the river were washed into the cave, gradually leveling the uneven bottom and creating a spacious, flat floor. It was at this point that Peking man took up residence.

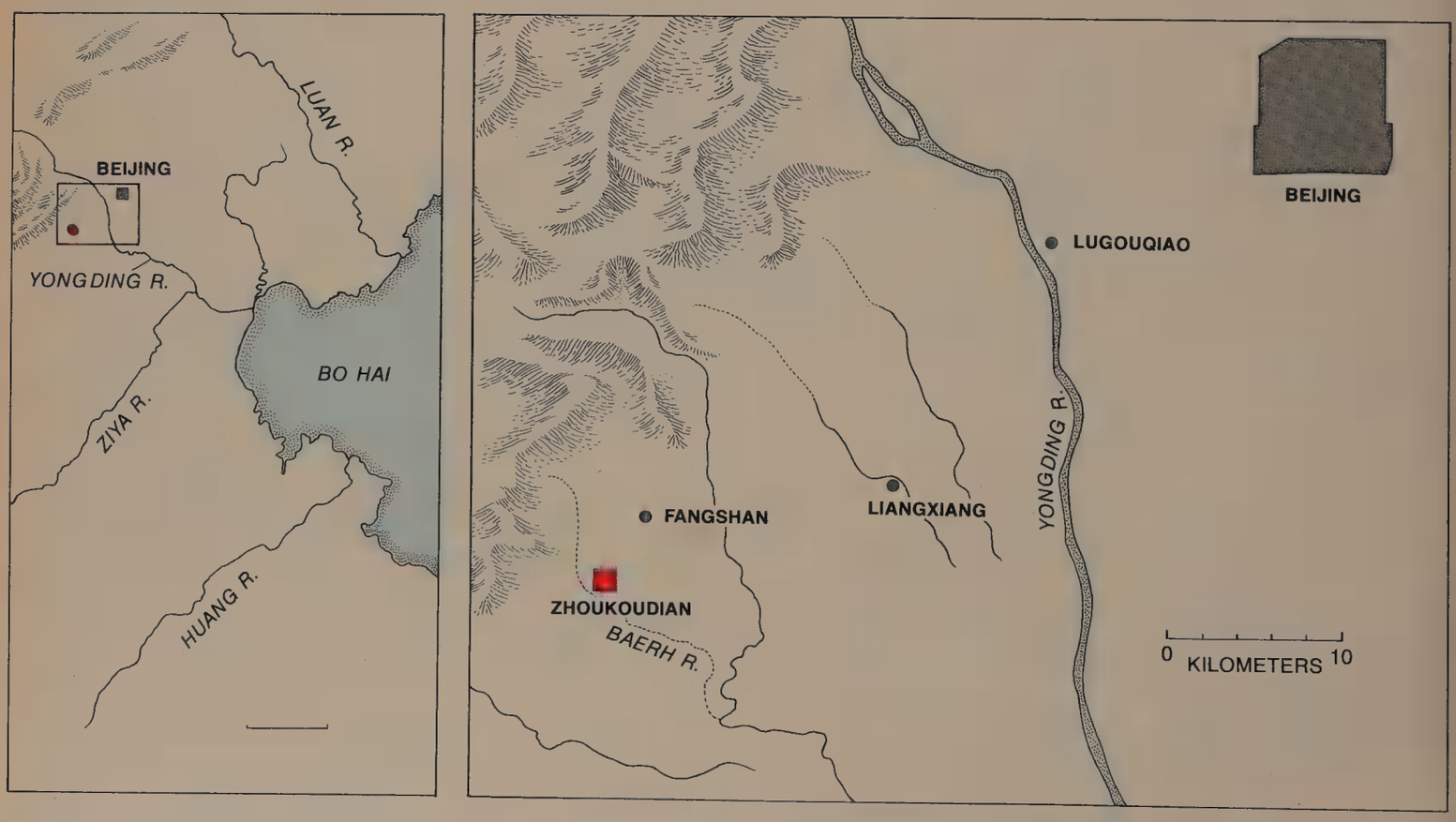
At first Peking man made his entrance from the east and lived mainly in the eastern section. About 350,000 years ago, however, there was a massive cave-in in the part of the eastern section now called the Dove Hall. An enormous amount of eroded rock fell from the roof, blocking the east entry completely. Peking man had to abandon what had been the most spacious part of the cave, but he was able to enter the cave again through a fissure in the central section. He thereafter lived in the western section until 230,000 years ago, when the cave was completely filled with the detritus of human occupation and with eroded limestone.

The cave remained filled with these deposits until its excavation. The deposits that included remains of Peking man had a thickness of about 40 meters, which was divided into 10 layers by Pierre Teilhard de Chardin and Yang Zhongjian, a Chinese paleontologist, in 1929. The Chinese archaeologist Jia Lanpo analyzed the deposits below the 10th layer in 1951 and defined three more layers (11 through 13). The various investigators have agreed to identify the deposits as those of the Zhoukoudian period. Although the period was generally recognized as belonging to the Middle Pleistocene, about half a million years ago, until recently the chronology was so sweeping and lacking in detail that it gave no idea about the evolution of Peking man himself.

Dozens of workers in five institutions have recently finished a series of chronological investigations and so for the first time have given a comparatively exact age for each layer. The 10th layer was dated by counting the tracks left in crystalline minerals by the disintegration of nuclei of uranium 238, the most abundant isotope of uranium. The rate of such fission events is constant, and so the number of tracks is a measure of the time that has elapsed since the mineral formed. It was concluded that the 10th layer was laid down 460,000 years ago.

Workers at the Institute of Geology of the Chinese Academy of Sciences have ascertained by the dating technique called the uranium-series method that the ninth and eighth layers were deposited 420,000 years ago. In the uranium-series method what is measured are the relative abundances of uranium 234 and the product of its radioactive decay, thorium 230. The decay proceeds at a constant, known rate; the ratio of the two elements therefore indicates the time since the uranium-bearing mineral was formed.

Still another method was applied to the dating of the seventh layer, which turns out to be between 370,000 and 400,000 years old. The age was established by examining the natural remanent magnetism of minerals in the layer. When a magnetic mineral crystallizes, its magnetic axis becomes aligned with the direction of the earth's mag-



SITE OF THE CAVE is the small town of Zhoukoudian, some 50 kilometers southwest of Beijing (*right*). It is administratively subordinate to Fangshan, a suburban county of Beijing about 150 kilometers from the seacoast (*left*). In the northwestern part of the town is Dragon-Bone Hill, named for the mammalian fossils collected there by local residents, who thought they were dragon bones of great medicinal value. Zhoukoudian Locality I, which is one of four paleoanthropologic sites on the hill, is the cave that was inhabited by Peking man.

netic field. Periodic reversals of the earth's field thus leave a record in the rocks, which identifies their time of deposition. The uranium-series method has also been applied to the deposits in the three topmost layers. The results indicate an absolute age of 230,000 years.

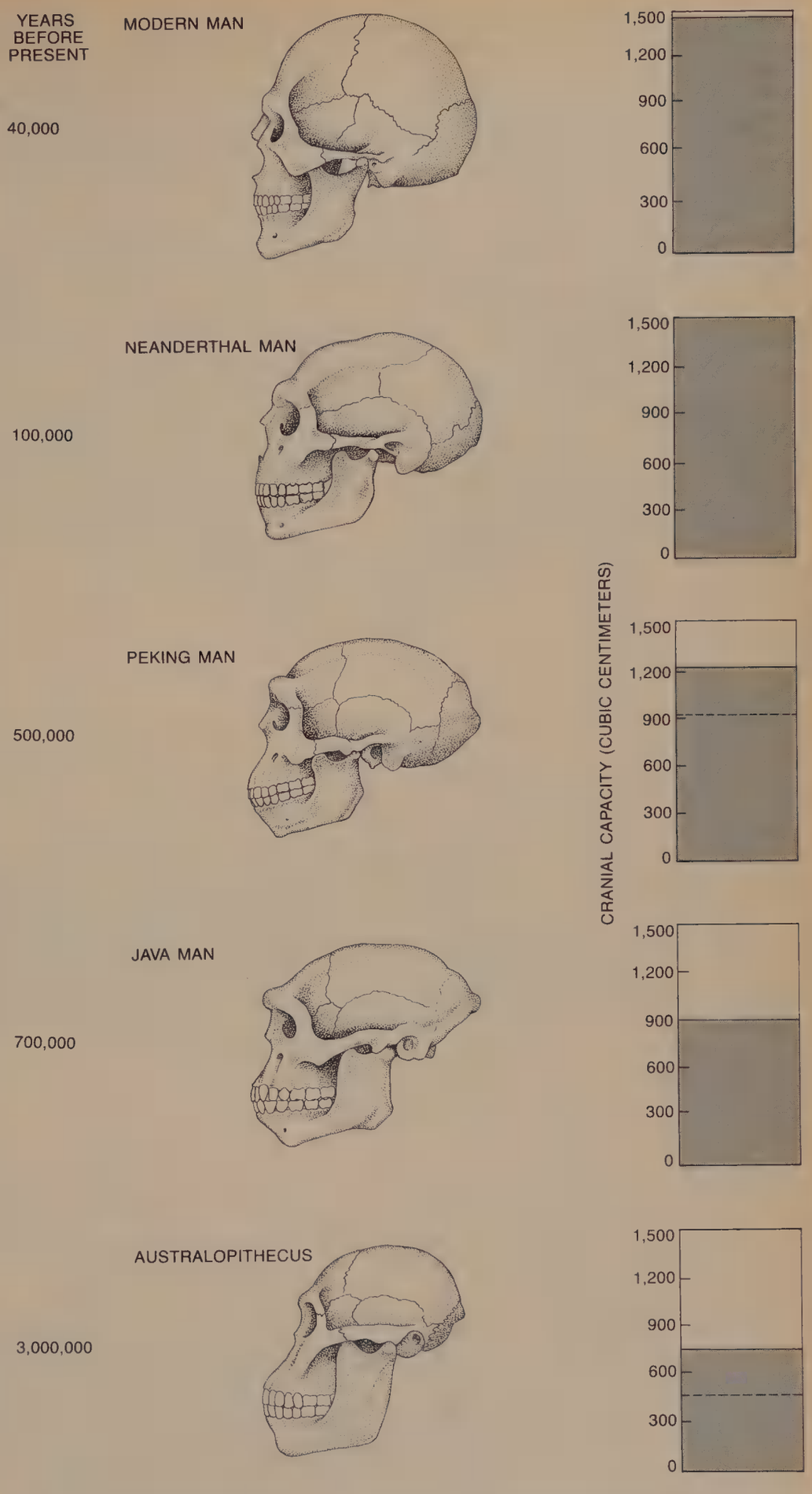
The Fossils of Peking Man

The fossils of Peking man discovered over the past 50 years have been catalogued as follows: six complete or almost complete skulls and 12 other skull fragments, 15 pieces of mandibles, 157 teeth, three fragments of humerus (the bone of the upper arm), one clavicle (collarbone), seven fragments of femur (the thighbone), one fragment of tibia (shinbone) and one lunate bone (a wristbone shaped like a half moon). They were found at 15 widely scattered places in the cave. The fossils represent material from more than 40 male and female individuals, who died at various ages. From the fossil material it is possible to reconstruct the physical form and appearance of Peking man.

Like other examples of *Homo erectus* that emerged in the Middle Pleistocene, Peking man had a skeleton much like that of modern man. The main difference is that the limb bones had a thicker wall and a smaller marrow cavity. The skull of Peking man, however, differed substantially from the modern form. It was much thicker and flatter and had protruding brows and a marked angle at the rear. The brain case was larger than that of *Homo habilis,* a species that flourished 1.8 million years ago, and even larger than that of Java man, the somewhat older form of *Homo erectus;* the cranial capacity was still much smaller than that of modern man. The teeth were larger and more robust than those of *Homo sapiens* and had traces of an enamel collar, called a cingulum, around the crown. The cingulum is a primitive trait in prehistoric man.

All fossils of Peking man found so far have been analyzed by one of us (Wu Rukang) and Dong Xingren of the Institute of Vertebrate Paleontology and Paleoanthropology. We have given particular attention to the question of Peking man's cranial capacity. It should be stressed that brain size cannot be taken unconditionally as an indicator of human intelligence; a hominid with a bigger brain case is not necessarily smarter. Nevertheless, modern man, with an average cranial capacity of 1,450 cubic centimeters, must have evolved from an earlier form of man with a smaller cranium. The tendency is therefore for the human cranial capacity to increase, and the capacity of a given species or subspecies may provide some information about its place in human evolution.

It is generally accepted that the cranial capacity of man and his ancestors was

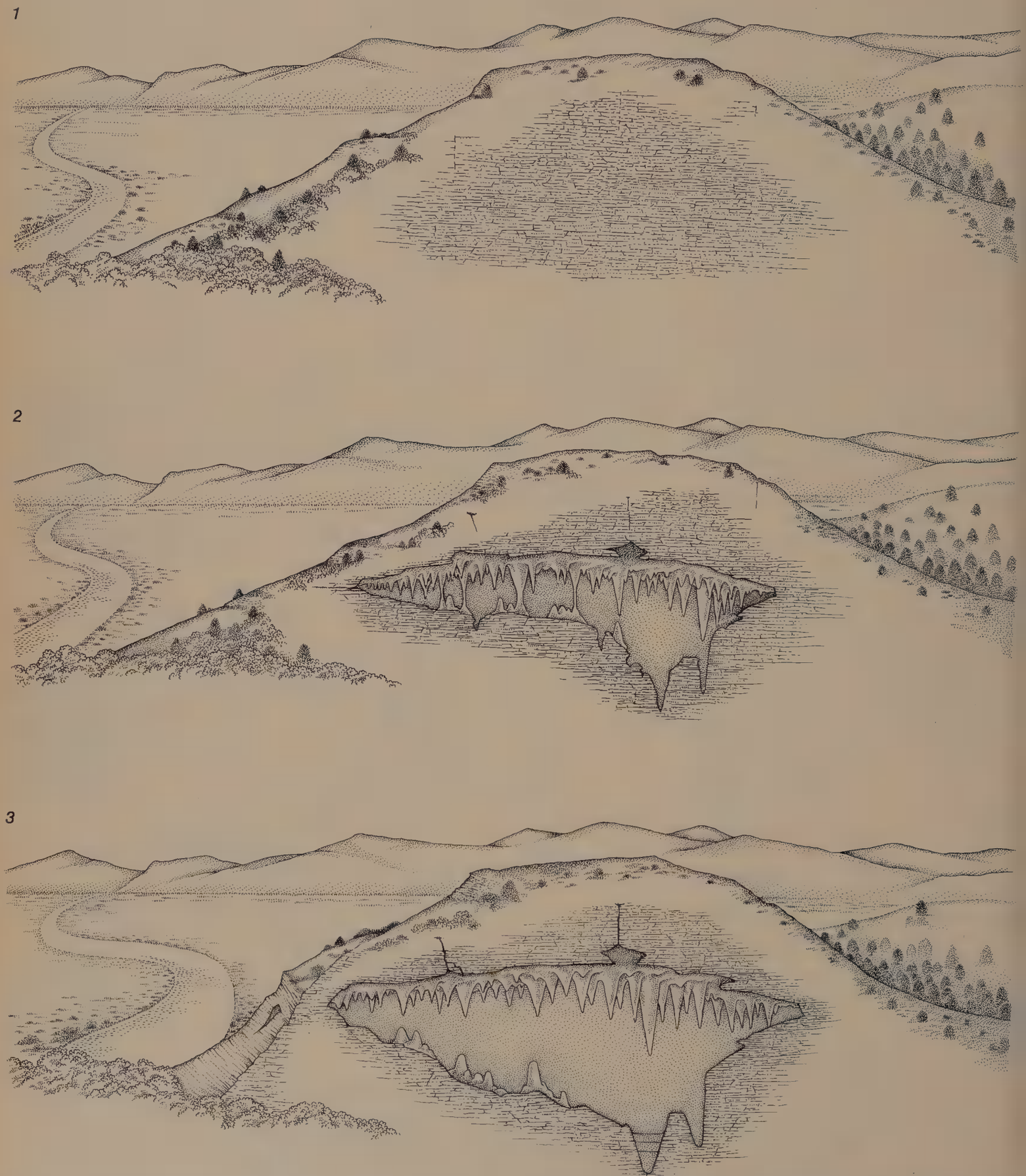


HUMAN EVOLUTION can be traced through morphological changes in the skull (*left*) and through an increase in cranial capacity (*right*). The early hominid *Australopithecus,* which appeared some four million years ago, had a flat skull with protruding jawbones and a small brain case (between 450 and 750 cubic centimeters). About 1.5 million years later came *Homo erectus.* Two examples of *Homo erectus,* Java man and Peking man, were comparatively advanced in their facial structure and cranial capacity, which ranged from 850 cubic centimeters to more than 1,000 cubic centimeters. They differed in their age, skull shape and brain size. Not until some 100,000 years ago did Neanderthal man, the first member of the species *Homo sapiens,* come into existence; by then the brain case was much larger and the jawbones were less protuberant. Modern man, the subspecies *Homo sapiens sapiens,* emerged about 40,000 years ago.

stable at from 500 to 800 cubic centimeters in the interval from three to 1.5 million years ago. This is the cranial size measured in fossils of *Australopithecus* and *Homo habilis*. Java man, the form of *Homo erectus* that appeared about a million years ago, had a cranial capacity of some 900 cubic centimeters. When Peking man emerged, the average capacity had reached 1,054 cubic centimeters.

It should be pointed out that this brain-case value for Peking man is an average based on the six relatively complete skulls found at Zhoukoudian. One of the six is the skull of a child who died at the age of eight or nine. A juvenile skull is obviously smaller than that of an adult; if it were not taken into account, the average cranial capacity of Peking man would be 1,088 cubic centimeters.

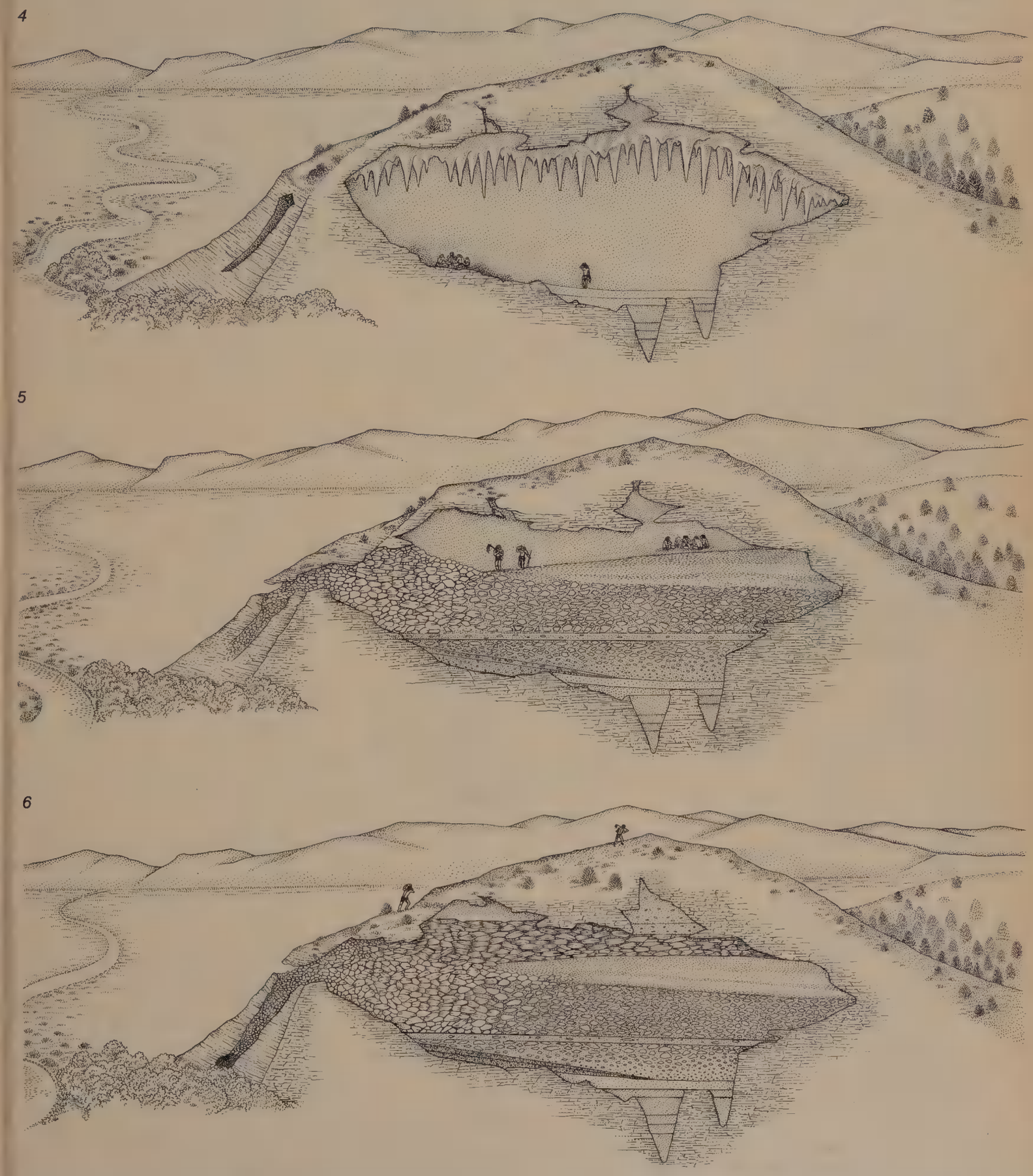
Because of the long time span covered



EVOLUTION OF THE CAVE at Zhoukoudian Locality I is schematized here in six stages. Dragon-Bone Hill, in which the cave was created, is a limestone structure formed about 450 million years ago (*1*). As limestone was dissolved by underground water the cave began to form inside the hill about five million years ago (*2*). When the eastern spur of the hill was cut off by the scouring of the ancient Zhoukou River some three million years ago, a small entry to the cave appeared on the eastern slope and became progressively larger (*3*). Sand and

by the fossils of Peking man, it may even be possible to trace changes in morphology, and in particular an increase in cranial capacity, over the history of the habitation of the site. The average capacity of four skulls found in the eighth and ninth layers (which are dated to more than 400,000 years ago) is 1,075 cubic centimeters. The one skull unearthed in the third layer (230,000 years old) has a capacity of 1,140 cubic centimeters. It seems that Peking man became more manlike in anatomy after 200,000 years of cave dwelling.

It is acknowledged in the anthropological community that the morphological evolution of early man was much slower and so less obvious than the transformation of behavior and life-style. Toolmaking techniques and the tools themselves are therefore an important



dirt were carried by the river into the cave, leveling the gullies created by the erosion of the limestone. The resulting spacious shelter was adopted by Peking man about 460,000 years ago (*4*). The east entry and the eastern part of the cave were the main areas of habitation until the collapse of the roof 300,000 years ago forced ■ move to the western part, with access through ■ fissure in the central section (*5*). By the time Peking man left the cave 230,000 years ago it had been filled with fallen rock and the detritus of human occupation (*6*).

measure of human evolution. In the case of Peking man a systematic study of the lithic culture was done three years ago by Pei Wenzhong and Zhang Shenshui of the Institute of Vertebrate Paleontology and Paleoanthropology. They came to believe Peking man's evolution is more clearly mirrored in his tools and toolmaking behavior than in his fossil remains.

The Tools of Peking Man

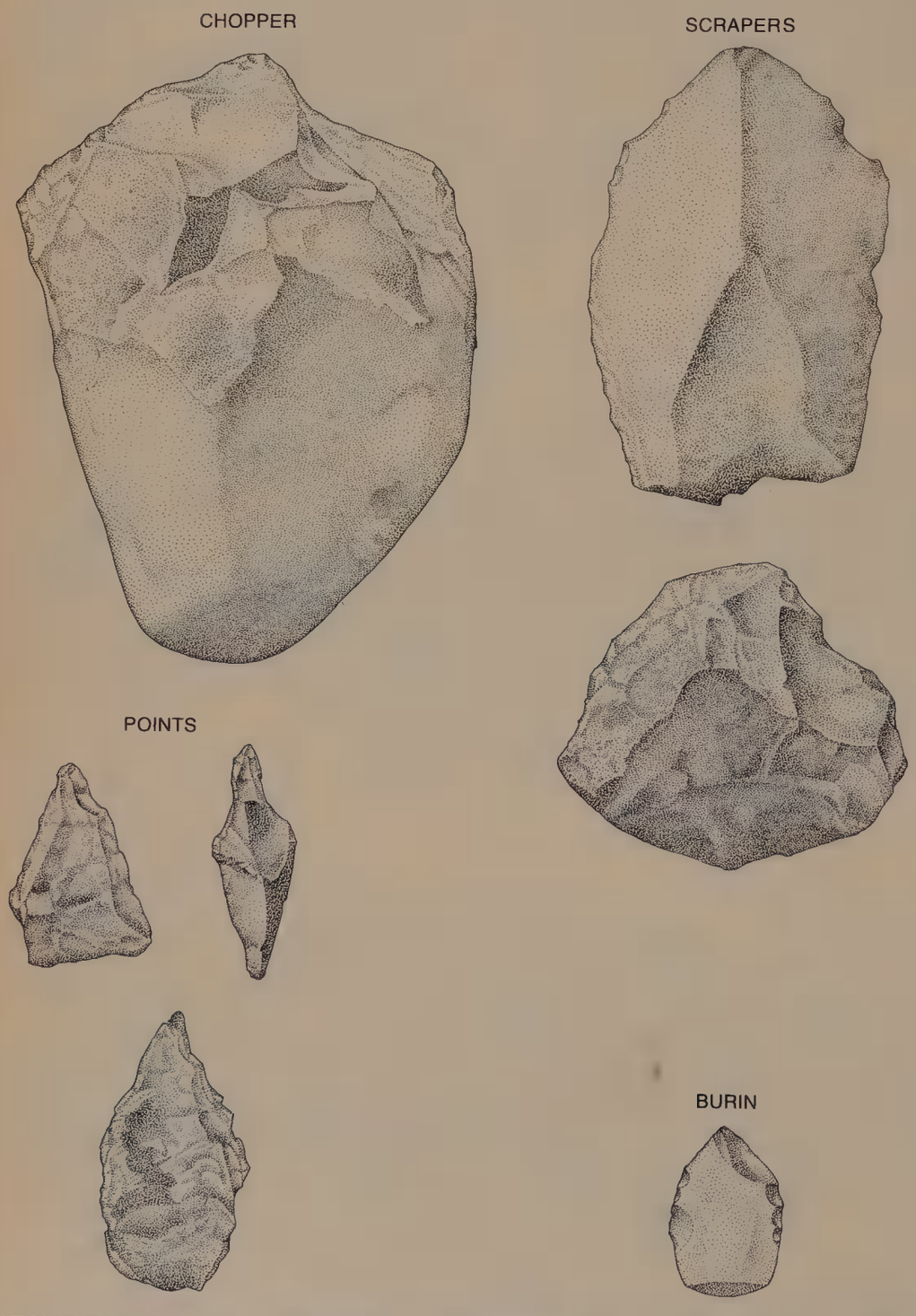
The stone industry of Peking man was an advanced one both in the selection of materials exploited and in the toolmaking techniques applied to them. The stone artifacts produced by Peking man are primarily made of vein quartz, rock crystal, flint and sandstone, suggesting that he did not rely exclusively on water-rounded pebbles for tool material. He often had to walk some distance to get vein quartz and rock crystal where they had been exposed by the weathering of granite formations.

Most of the tools are flakes of various sizes made in one of three ways. In the technique called anvil percussion a large flat stone (the anvil) was placed on the ground and forcefully struck with a piece of sandstone. Flakes chipped off the sandstone were gathered up, and those with a suitable shape and edge were selected either for direct use or for later retouching. In another method, called direct percussion, the core flint was held in one hand and flakes were detached from it by striking it with a hammerstone held in the other hand. The third flaking method is called bipolar percussion. Again a large, flat anvil was set on the ground. A piece of vein quartz was then held upright on the anvil with one hand and crushed by vertical blows with a hammerstone grasped in the other. In this way flakes could be chipped off both ends; they are called bipolar flakes.

Although some of the flakes recovered at Zhoukoudian were used as they were, it is evident that some of them had been trimmed into more specialized tools such as scrapers, points, choppers, burins and awls. The retouching was done mainly by the direct-percussion process. Most of the tools were trimmed on their back surface rather than on the one that had been worked in the initial flaking operation. Sometimes both faces were retouched.

The evolution of Peking man's lithic culture was divided into three periods by Pei and Zhang. The division was made according to changes in tool size, in tool material and in toolmaking technique. The stone tools found in the eighth through the 11th layers represent the earliest culture, which prevailed between 460,000 and 420,000 years ago. It is characterized by large tools weighing more than 50 grams and longer than 60 millimeters and by the indiscriminate application of the three toolmaking processes. At this stage tools made out of softer materials such as sandstone account for 15 to 20 percent of the artifacts. In the middle stage, from 370,000 to 350,000 years ago, the anvil-percussion process was practically abandoned and the bipolar-percussion process became the main method of flaking stone. As a result the proportion of the tools having a weight of less than 20 grams and a length of less than 40 millimeters increased to 68 percent. In contrast large tools diminished to 12 percent.

The last stage lasted from 300,000 to 230,000 years ago. It was clearly the most advanced: the tools had become smaller and the tool materials were of better quality. Among the stone tools excavated from layers one through five the small tools increased dramatically to 78 percent of the total, whereas the large tools decreased further to 5 percent. Although the tools of the third period were still made mainly out of quartz, fewer were made out of the coarse varieties of the stone, such as vein quartz, and the fraction of the tools made out of flint increased to as much as 30 percent in the uppermost layers. In the meantime sandstone tools



STONE TOOLS made by Peking man are of four main types: the chopper (*upper left*), the scraper (*upper right*), the point (*lower left*) and the burin (*lower right*). They were formed out of vein quartz, rock crystal, flint and sandstone. Some of the flakes were put into service as tools without further trimming; others were retouched into more specialized tools such as points and burins. The stone industry was dominated by small tools with a length of less than 40 millimeters and a weight of less than 20 grams, although most of the early tools are larger.

USE OF FIRE by Peking man, attested to by layers of ashes, can be traced to the beginning of his residence in the cave. Four layers of ashes have been identified in the cave deposits; the thickest layer is some six meters deep and the thinnest one is a little more than a meter deep.

had diminished to less than 1 percent of the total.

Another indication of the advanced state of Peking man's culture is his use of fire: evidently the art had already been mastered at the very beginning of his habitation of the cave. There are four large, thick layers of ashes deposited in periods ranging in age from 460,000 to 230,000 years. The thickest one, formed between 310,000 and 290,000 years ago, is six meters deep in certain places. Some of the ash deposits are in scattered piles, suggesting that Peking man had the ability to control fire and keep it burning for a long time.

How did Peking man make a fire and control it? A conclusive answer cannot be given, but an inference can be drawn. Peking man seems to have been too primitive to start a fire; instead he was probably dependent on natural fires outside the cave, generally started by lightning. Presumably the fire was captured by lighting a bunch of twigs or some other kindling and bringing it back to the cave. Given the rarity of natural fires, it must have been vitally important to avoid letting the fire go out. One way of maintaining a fire would have been to add wood to it continuously. Another would have been to cover burning charcoal with ash or soil to slow its burning. A new fire could then be made by blowing on the coals. Some charcoal found in the cave may be a remnant of the latter process.

What was the nature of the environment outside the cave at the time Peking man occupied the site? On the basis of a study of pollen at Zhoukoudian Locality I, Kong Zhaochen of the Institute of Botany of the Chinese Academy of Sciences and his colleagues have suggested that Peking man lived in the climate of an interglacial period; indeed, the climate of the period was not much different from that of northern China today. The vegetation consisted of temperate deciduous forests and steppes on the plains and in the valleys and coniferous forests on the mountains.

A temperate climate entails the rigors of a cold winter every year but also provides many kinds of edible plant products. The adoption of a cave as shelter and the use of fire may have been inspired directly by the need for surviving cold weather.

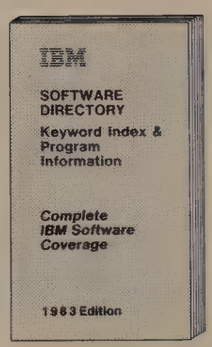
A variety of plants supplied not only firewood but also fruits and seeds. Among the deposits in the cave are quantities of charred seeds of the Chinese hackberry, which were obviously gathered and roasted by Peking man. Evidently the seeds served as food. The analysis of pollen in the deposits gives evidence of other plants, such as walnut, hazelnut, pine, elm and rambler rose. The fruits and seeds of those species must also be considered candidates for inclusion in Peking man's diet.

Hunting was a valuable adaptation to the environment since meat could supply more calories and protein than a vegetarian diet. Peking man was evidently able to compete successfully with large carnivores as a hunter. An abundance of fossil bones of mammals of various sizes found in the cave indicates that Peking man not only hunted small game but also was capable of killing large animals.

Among carnivores a natural relation

Get IBM's free 1983 software directory!

The brand-new 1983 edition of the IBM Software Directory lists key facts for over 2,000 software programs and aids. These programs cover the entire range of IBM systems, including the IBM System/370, 4300, 30xx, 8100, 51xx, Systems 3, 32, 34, 38, System/23-Datamaster, Series/1, the IBM Displaywriter and the IBM Personal Computer.



Get your free copy. Fast.

Call 1 800 631-5582 or send us the coupon now.

IBM®

SA 6/83

IBM Direct
1 Culver Road
Dayton, NJ 08810

Please rush my free copy of the IBM Software Directory.

Name

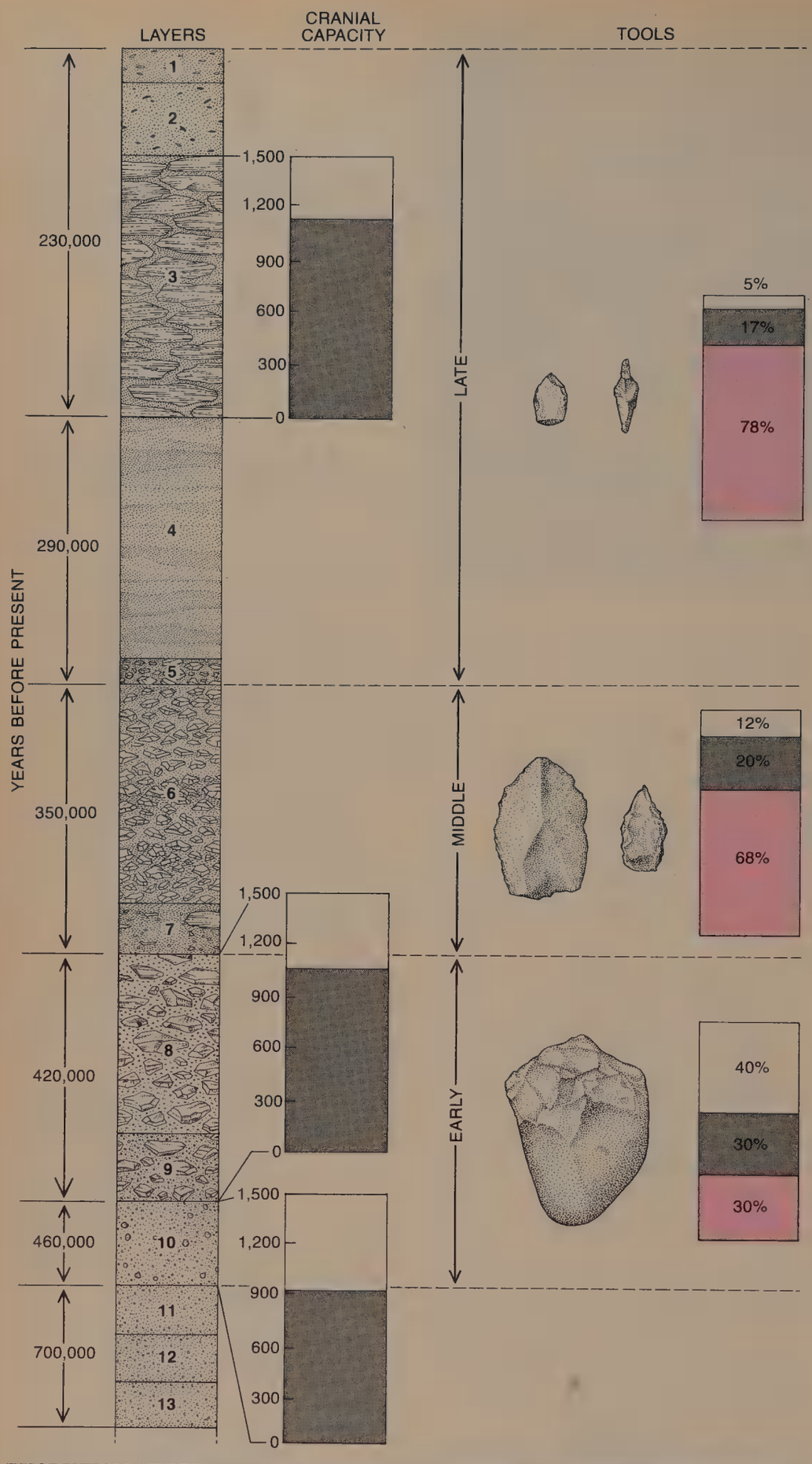
Title

Company

Address

City State

Zip Phone



EVOLUTION OF PEKING MAN in the course of 230,000 years of cave dwelling is suggested by an increase in brain size (*left*) and the development of a more refined lithic culture (*right*). The cave deposits that include fossils of Peking man are divided into 13 layers. Skulls have been recovered from layer 10, from the boundary between layer 9 and layer 8 and from layer 3. The measured cranial capacities are 915 cubic centimeters for the earliest skull, an average of 1,075 cubic centimeters for four later skulls and 1,140 cubic centimeters for the most recent one. It seems the brain size increased by more than 100 cubic centimeters during the occupation of the cave. The cultural development of Peking man is broken down into three stages. In the earliest stage the artifacts were mainly choppers and scrapers; 40 percent were large tools (*white*), 30 percent were of intermediate size (*gray*) and 30 percent were small (*color*). In the latest stage more complex tools appeared. The proportion of large and medium tools decreased respectively to 5 and 17 percent, whereas the proportion of small ones increased to 78 percent.

generally exists between the size of the predator and that of its prey. A fox cannot kill a zebra, but the zebra is the favorite food of lions. Owing to his use of weapons, Peking man escaped the limitations of his size. In particular he became an efficient hunter of deer, some of which were larger and faster than he was. The large quantity of fossils belonging to at least 3,000 individual deer of two species, the thick-jawed deer (*Megaceros pachyosteus*) and the sika deer (*Pseudaxis grayi*), could be regarded as an indication that deer were the commonest prey of Peking man.

Social Adaptations

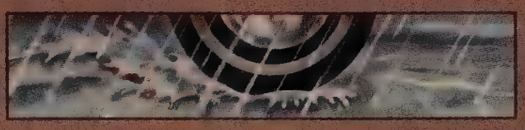
As we have shown, Peking man was a cave dweller, a fire user, a deer hunter, a seed gatherer and a maker of specialized tools. The fossil and artifactual evidence attest to his biological and technological adaptation. It is more difficult to trace his social adaptation since little evidence is available. Nevertheless, three hypotheses can be put forward on the basis of our studies.

Generally speaking, gathering is a simple labor that can be undertaken by single individuals. Hunting, in contrast, and particularly the hunting of large animals, is so complicated, difficult and hazardous that the cooperation of numerous individuals is needed. It can be inferred, therefore, that Peking man was more likely to have been living in a group than in solitude when he began to hunt deer. Furthermore, thousands of fossils of prey species found in the cave suggest that these primitive hunters may have preferred to bring prey back to the cave and share the meat with others in the community rather than to consume it where it was killed.

Second, the hunting of large and fast-moving animals such as deer may have been difficult for women because of physiological limitations (such as pregnancy and child rearing). It is thus possible to speculate that the hunting behavior of Peking man may have caused or contributed to the sexual division of labor within the group. The pattern of male hunters and female gatherers, which is common in hunting-and-gathering societies today, may have already been established.

Third, the existence of a consistently progressing lithic culture throughout a period lasting for 200,000 years or more suggests that the earliest practice of education may have taken place in Peking man's cave. It is out of the question that each generation could recapitulate the entire history of stone-tool development, from striking a pebble in order to make a crude tool to trimming a flake into a specialized scraper. Tool-making techniques, like modern science and technology, must have been conveyed from the old to the young.

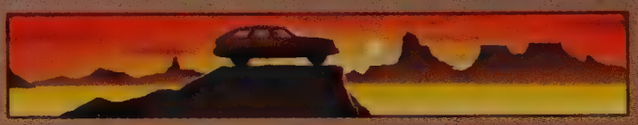
FOR THE WET



FOR THE WHITE

FOR THE WINDING

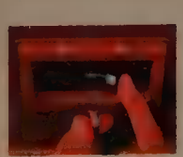
FOR THE WILD



FOR THE WARM

FOR THE WEEKEND

FOR THE WALLET



W is for Eagle Wagon. And for the wherever. Because Eagle is the only automobile in the world that lets you go from 2-wheel drive to full-time 4-wheel drive at the flick of a switch. 2-wheel drive for high mileage on the highway. 4-wheel drive for worry-free traction and security when the roads weave, the weather worsens, or when you just want to wander where there is no road.

And Eagle's well-appointed interior makes the going all the more wonderful. Why not see the 1983 Eagle Wagon today. Wherever you're going, whatever the weather, it'll fit you to a **W.**

The 2-wheel/4-wheel drive EAGLE

FROM AMERICAN MOTORS

*Optional 5-speed stick. Use these figures for comparison. Your results may differ due to driving speed, weather conditions and trip length. Highway and California figures lower.

Dark Matter in Spiral Galaxies

It appears that much of the matter in spiral galaxies emits no light. Moreover, it is not concentrated near the center of the galaxies

by Vera C. Rubin

After evidence was obtained (in the 1920's) that the universe is expanding it became reasonable to ask: Will the universe continue to expand indefinitely or is there enough mass in it for the mutual attraction of its constituents to retard the expansion and finally bring it to a halt? Most cosmologists agree that the universe started in a big bang 10 to 20 billion years ago from an infinitely small and dense state and that it has been expanding ever since. It can be calculated that the critical density of matter needed to brake the expansion and "close" the universe is on the order of 5×10^{-30} gram per cubic centimeter, which is equal to about three hydrogen atoms per cubic meter. The amount of luminous matter in the form of galaxies, however, comes to only about 7.5×10^{-32} gram per cubic centimeter. Therefore if the expansion of the universe is to stop, the density of the invisible matter must exceed the density of the luminous matter by a factor of roughly 70.

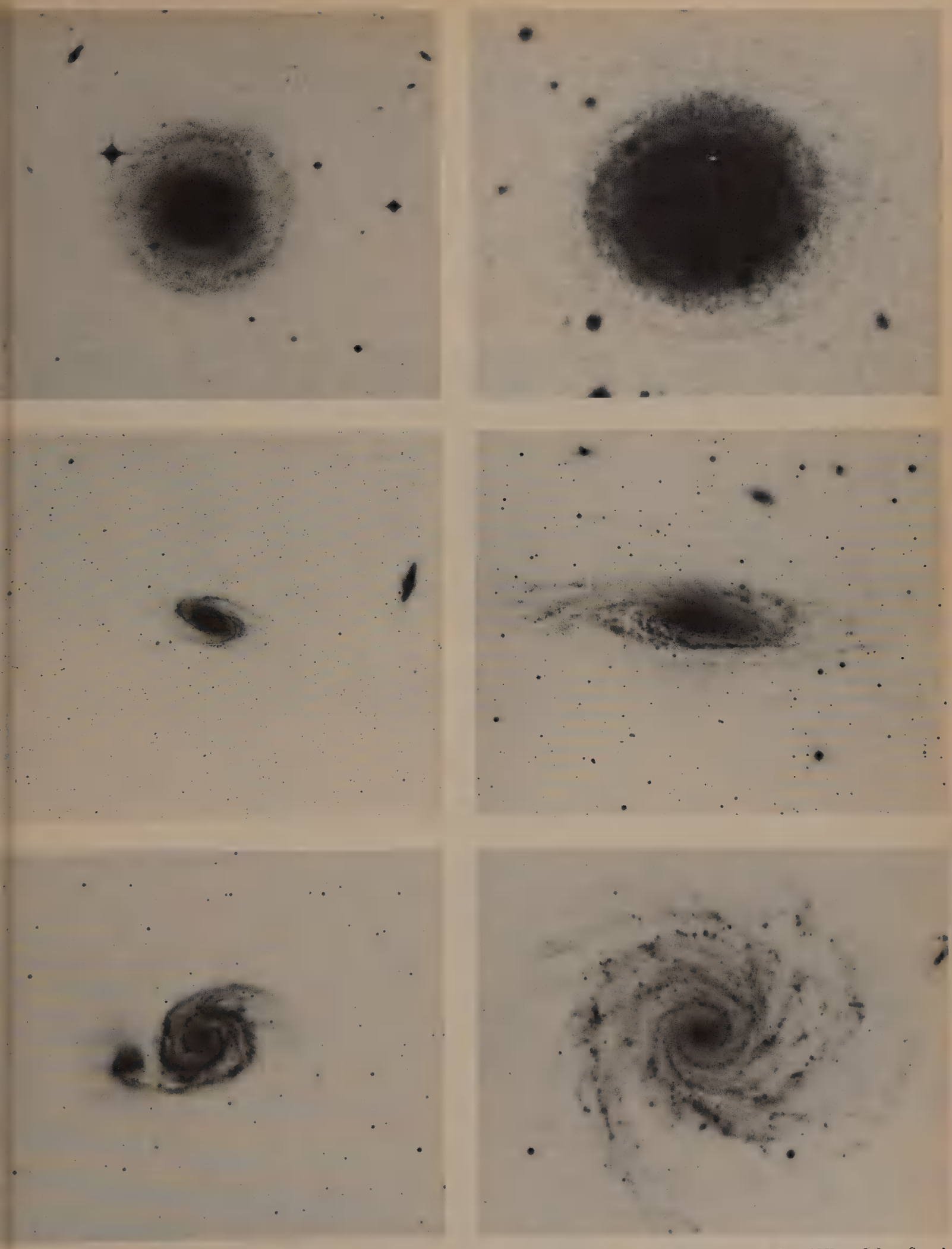
With this factor in mind astronomers over the past half century have sought to determine the mass of the galaxies that populate the universe out to the limits of observation. From the luminosity of typical galaxies one can estimate that they have a mass ranging from a few billion to a few trillion times the mass of the sun. The actual stellar population of a galaxy is of course highly diverse. Some stars are 10,000 times more luminous than the sun per unit of mass; others are only a small fraction as luminous. Given this diversity one would like to know: Is the distribution of luminosity in galaxies a reliable indicator of the distribution of mass? And, by extrapolation, is the distribution of luminosity in galaxies a reliable indicator of the distribution of mass in the universe?

My colleagues and I in the Department of Terrestrial Magnetism of the Carnegie Institution of Washington have sought to answer these questions by measuring the rotational velocity of selected galaxies at various distances from their center of rotation. It has been known for a long time that outside the bright nucleus of a typical spiral galaxy the luminosity of the galaxy falls off rapidly with distance from the center. If luminosity were a true indicator of mass, most of the mass would be concentrated toward the center. Outside the nucleus the rotational velocity would fall off inversely as the square root of the distance, in conformity with Kepler's law for the orbital velocity of bodies in the solar system. Instead it has been found that the rotational velocity of spiral galaxies in a diverse sample either remains constant with increasing distance from the center or rises slightly out as far as it is possible to make measurements. This unexpected result indicates that the falloff in luminous mass with distance from the center is balanced by an increase in nonluminous mass.

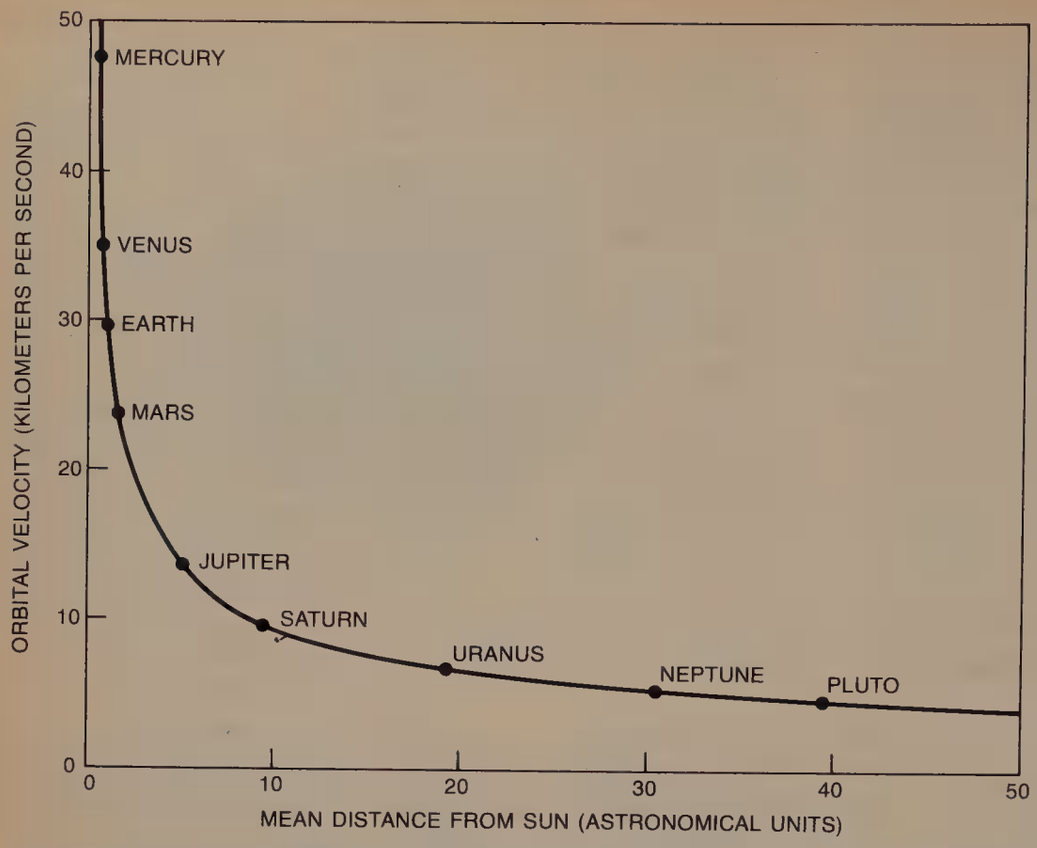
Our results, taken together with those of many other workers who have attacked the mass question in other ways, now makes it possible to say with some confidence that the distribution of light is not a valid indicator of the distribution of mass either in galaxies or in the universe at large. As much as 90 percent of the mass of the universe is evidently not radiating at any wavelength with enough intensity to be detected on the earth. Originally astronomers described the nonluminous component as "missing matter." Today they recognize that it is not missing; it is just not visible. Such dark matter could be in the form of extremely dim stars of low mass, of large planets like Jupiter or of black holes, either small or massive. Other candidates include neutrinos (if indeed they have mass, as recent work suggests) or such hypothetical particles as magnetic monopoles or gravitinos.

Early in this century it was reasona-

SPIRAL GALAXIES whose unseen mass has been investigated by the author fall into three principal classes: Sa, Sb and Sc. Within each



class the galaxies vary enormously in size and luminosity. Here nine examples are reproduced in negative images in which 1.2 centimeters equals 10 kiloparsecs (32,600 light-years). Three Sa galaxies are at the top, three Sb galaxies are in the middle and three Sc galaxies are at the bottom. In the progression from Sa to Sc the nucleus gets smaller with respect to the disk and spiral structure gets more pronounced.



KEPLER'S LAW for the orbital velocity of planets in the solar system, in which more than 99 percent of the total mass resides in the sun, yields this plotted curve. Orbital velocity decreases inversely as the square root of *r*, the planet's mean distance from the sun. The distance is shown here in astronomical units; one A.U. equals the mean distance between the earth and the sun. Pluto, at 39.5 A.U., lies 100 times farther from the sun than Mercury, at .39 A.U. Mercury's orbital velocity is about 47.9 kilometers per second; Pluto's velocity is accordingly slower by a factor of 10, or 4.7 kilometers per second ($47.9 \times 1/\sqrt{100}$). The author's results show that the orbital velocities of stars in a spiral galaxy depart strongly from a Keplerian distribution.

ble for astronomers to assume that the distribution of luminous matter, wherever it was found, coincided with the distribution of mass. Nearly 50 years ago, however, Sinclair Smith and Fritz Zwicky of the California Institute of Technology discovered that in some large clusters of galaxies the individual members are moving so rapidly that their mutual gravitational attraction is insufficient to keep the clusters from flying apart. Either such clusters should be dissolving or there must be enough dark matter present to hold them together. Almost all the evidence suggests that clusters of galaxies are stable configurations. Hence the early observations of Smith and Zwicky marshaled the first evidence that such clusters harbor matter both luminous and nonluminous.

Recent work by many other astronomers has strengthened this conclusion. Studies of the dynamics of individual galaxies, including our own galaxy, of pairs of galaxies, of groups of galaxies and of clusters of galaxies all point to a component of unobservable but ubiquitous mass. Such studies detect the presence of nonluminous mass solely by its gravitational effects.

For the past several years W. Kent Ford, Jr., Norbert Thonnard, David Burstein and I have sought to learn about the distribution of mass in the universe by investigating the distribution of matter within galaxies with a structure similar to that of our own galaxy, namely the general class of spiral galaxies. We have adopted this approach because spiral galaxies have a geometry favorable for the identification of mass, whether it is luminous or nonluminous, and modern large telescopes equipped with image-tube spectrographs make it possible to complete an observation of a single galaxy with an exposure of about three hours. Before I describe our observations it will be helpful if I review how celestial objects respond to the gravitational force acting on them and how that response can reveal the large-scale distribution of matter.

Toward the end of the 17th century Robert Hooke suspected that the planets were subject to a gravitational force from the sun whose intensity decreased inversely as the square of the distance. Isaac Newton then recognized that all pairs of objects in the universe have a gravitational attraction for each other that is proportional to the product of their masses and inversely proportional to the square of the distance between them. In other words, if the distance between the objects is increased by, say, a factor of two, their mutual attraction decreases by a factor of four.

For planets in orbit around the sun, which embodies essentially all the mass in the solar system, the decrease in gravitational attraction with distance is exactly paralleled by a decrease in the velocity needed to hold the planet in its orbit. Therefore Mercury, lying at .39 astronomical unit from the sun (that is, .39 of the mean distance between the sun and the earth), has an orbital velocity of about 47.9 kilometers per second. Pluto, 100 times farther away at a mean distance of 39.5 astronomical units, has an orbital velocity only a tenth that of Mercury, or 4.7 kilometers per second. Spiral galaxies rotate because they retain the angular momentum and the orbital momentum of the initial clumps of gas from which they formed.

In a spiral galaxy the gas, dust and stars in the disk of the galaxy (together with any associated planets and their satellites) are all in orbit around a common center. Like the planets in the solar system, the gas and stars move in response to the combined gravitational attraction of all the other mass. If the galaxy is visualized as a spheroid, the gravitational attraction due to the mass M_r lying between the center and an object of mass m in an equatorial orbit at a distance r from the center is given by Newton's law GmM_r/r^2, where G is the constant of gravitation. If the galaxy is neither contracting nor expanding, the gravitational force is exactly equal to the centrifugal force on the mass at distance r: $GmM_r/r^2 = mV_r^2/r$, where V_r is the orbital velocity.

When this equation is solved for V_r, the value of m drops out and the velocity of a body at distance r from the center is determined only by the mass M_r inward from its position. If, as in the solar system, virtually all the mass is near the center, then the velocities outward from the center decrease as $1/r^2$. Such a decrease in orbital velocity is called Keplerian after Johannes Kepler, who first stated the laws of planetary motion.

In a galaxy the brightness is strongly peaked near the center and falls off rapidly with distance. Astronomers had long assumed that the mass too decreased rapidly with distance, in accordance with the distribution of luminosity. Hence it was expected that stars at increasing distances from the center would have decreasing Keplerian orbital velocities. Until recently few velocity observations had been made in the faint outer regions of galaxies, either to confirm this expectation or refute it.

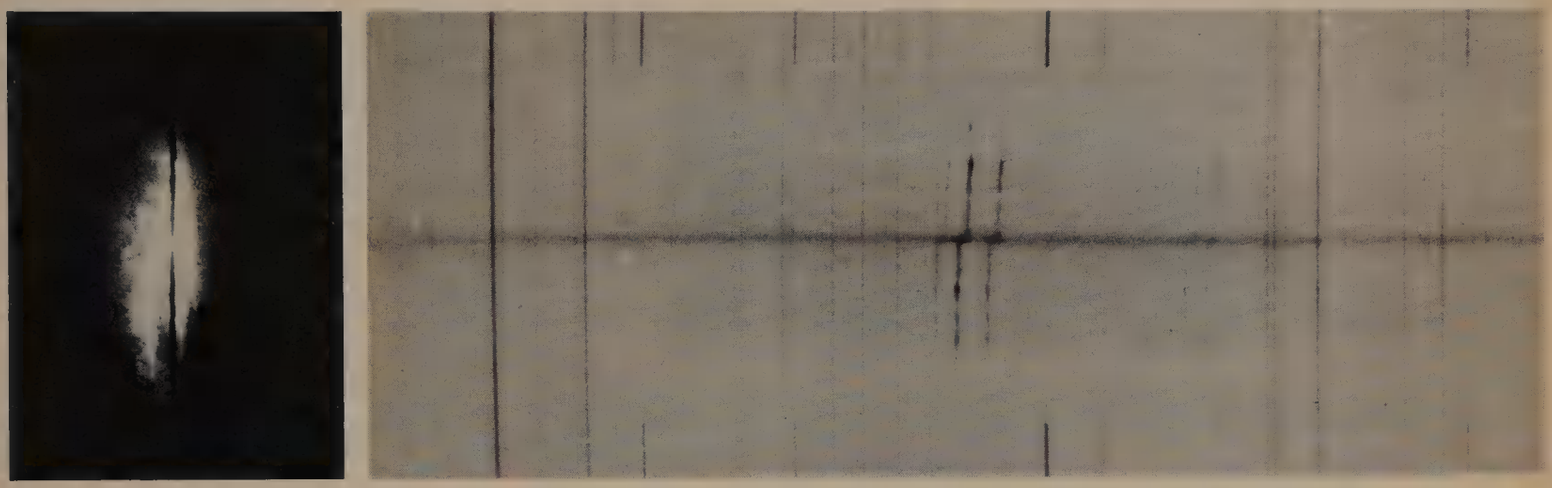
Although the forms of spiral galaxies are exceedingly diverse, astronomers are able to group them into three useful classes following a scheme proposed some 60 years ago by Edwin P. Hubble. Galaxies designated Sa have a large central bulge surrounded by tightly wound smooth arms in which "knots," or bright regions, are barely resolved. Sb galaxies have a less pronounced central bulge and more open arms with more pro-

nounced knots. Sc galaxies have a small central bulge and well-separated arms speckled with distinct luminous segments. The progression from type Sa to type Sc is one of decreasing prominence of the central bulge and increasing prominence of the disk rotating about it. That the disk is indeed rotating is assumed on simple dynamical grounds.

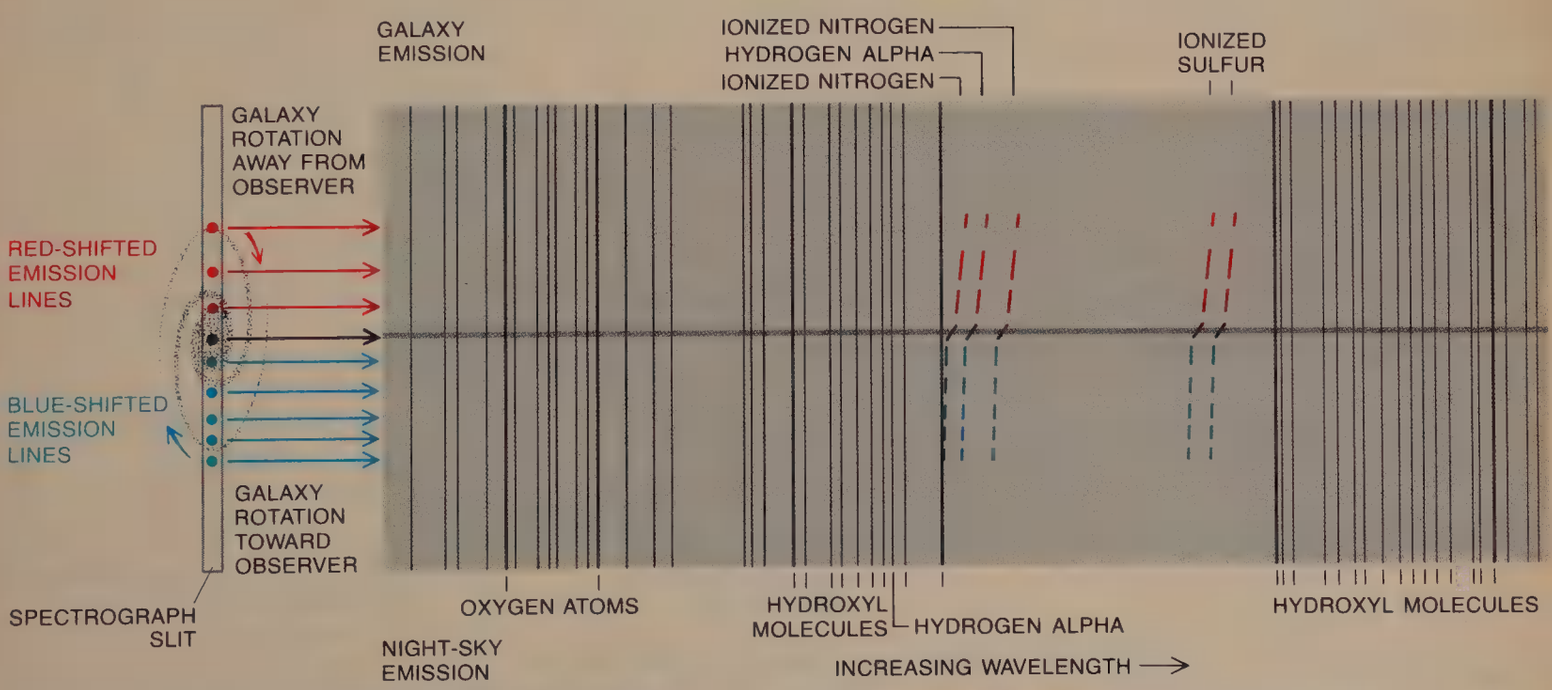
Within each type there are systematic variations in size and luminosity. For example, Sc galaxies range from small, low-luminosity, low-mass objects to galaxies of enormous luminosity and mass. For completeness, therefore, the study of the dynamics of galaxies should include not only objects with a range of morphological types but also objects with a range of luminosities.

Only for the closest stars in our own galaxy is it possible to detect motion by observing the changing position of the star against the background of more distant stars and galaxies on the celestial sphere. Even for the Andromeda galaxy, the large spiral galaxy closest to our own, it would take some 20,000 years for an orbital velocity of 200 kilometers per second (a velocity comparable to the sun's) to carry a star one second of arc across the sky. This is the minimum angular separation that can be detected optically from the earth. To study the motions in galaxies a different method is needed, one based on the phenomenon of the Doppler shift.

Doppler shifts are shifts in the frequency of waves from a source caused by the motion of the source toward or away from the observer. When the spectrum of the bright nucleus of a spiral galaxy is recorded, the absorption lines arising from the constituent stars are shifted toward the long-wavelength (red) end of the spectrum compared with the same lines in spectra made in laboratories on the earth. Such red shifts in the spectra of all but a few of the nearest galaxies, first observed in about 1915 by V. M. Slipher of the Lowell



SPECTRUM OF SPIRAL GALAXY NGC 7541 (*right*) was recorded with the four-meter telescope at the Kitt Peak National Observatory by the author and W. Kent Ford, Jr. NGC 7541 is a type Sc spiral, 60 megaparsecs distant. (A megaparsec is 3.26 million light-years.) The exposure time was 114 minutes. The galaxy is seen at the left as it appears on a television monitor in the telescope's console room. The dark line through the galaxy shows the orientation of the spectrograph slit. Light from across the disk is sampled (*see illustration below*).

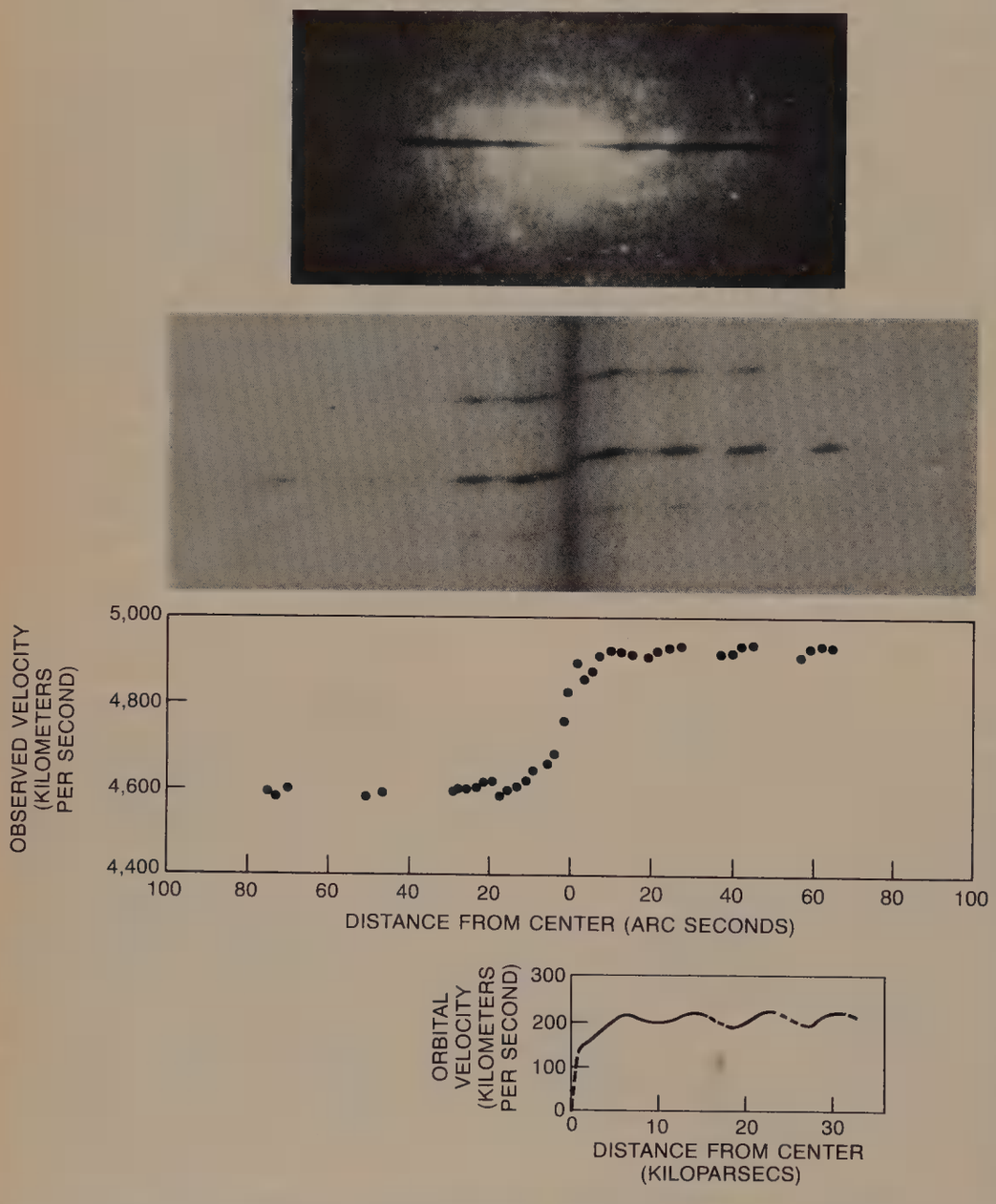


EMISSION LINES in the spectrogram of NGC 7541 arise from two sources: the night sky and atoms in the gas clouds surrounding bright stars in the galaxy. Most of the night-sky lines, which extend across the entire width of the spectrogram, are from hydroxyl (OH) molecules in the atmosphere of the earth. A few arise from oxygen and hydrogen atoms in the earth's atmosphere. The rotation of NGC 7541 shifts the position of the emission lines from the disk of the galaxy to either a shorter (bluer) wavelength or a longer (redder) one, depending on whether the rotation is carrying the stars and gas in the disk toward or away from the observer. Because the galaxy itself is traveling away from the observer as part of the general expansion of the universe, the hydrogen-alpha line from gas in the galaxy is red-shifted from the position of the same line in the night sky. The displacement is a measure of the galaxy's velocity of recession. The slant of the galactic emission lines shows that the orbital velocity of stars and gas in the disk is increasing with distance from the galactic center.

Observatory, provide the evidence that the universe is expanding, carrying almost all the other isolated galaxies away from ours and away from one another. As a result of Smith and Zwicky's work it is known that in pairs, groups and clusters of galaxies the local gravitational field overcomes the general expansion, so that these denser agglomerations of matter remain bound. Although the distances between clusters of galaxies are increasing, the distances between galaxies within clusters remain about the same. Slipher also noted that the spectra of individual galaxies can yield additional information about the motions of stars and gas within the galaxy.

If the disk of a spiral galaxy is oriented so that its plane is sharply tilted with respect to the line of sight from the earth, the rotation of the galaxy will carry the stars and gas on one side of the galactic nucleus toward our galaxy and those on the other side away from it. The spectral lines of the approaching material will therefore be blue-shifted, or raised in frequency, and the lines of the receding material will be red-shifted, or lowered in frequency. A measurement at any point on a spectral line will therefore supply both the angular distance of that point from the galactic nucleus and the velocity along the line of sight at that distance.

It is difficult to make spectroscopic measurements of the velocities of individual stars, which are faint even in galaxies fairly close to our own. In our work, therefore, we observe not stars but the light from the clouds of gas, rich in hydrogen and helium, that surround certain hot stars. The spectra of such clouds consist of bright emission lines that arise as an electron in an excited atom drops from a higher energy state to a lower one. In addition to emission lines of hydrogen and helium there usually are bright lines from atoms of nitrogen and sulfur that are singly ionized, or stripped of one electron. These lines are called forbidden because they arise only from atoms in the near-vacuum of space; in terrestrial laboratories such singly ionized atoms are rapidly deexcited by collisions with other atoms before the forbidden transition can occur.



MEASUREMENT OF THE ROTATION of NGC 2998, an Sc galaxy at a distance of 96 megaparsecs in the constellation Ursa Major, begins with the making of a spectrogram. The picture at the top shows the galaxy and superposed spectrograph slit as they appear on a television monitor at the four-meter Kitt Peak telescope. Below it is the hydrogen-alpha region of the spectrogram that resulted from an exposure of 200 minutes. The plotted points depict velocities across the galactic disk as measured from the hydrogen-alpha line. The entire galaxy is receding at 4,800 kilometers per second; the left side of the galaxy is approaching, the right side receding. The last step is to draw a rotation curve by smoothing velocities from both sides of the disk and translating angular distance on the sky into linear distance in the galaxy.

Until recently it was not possible to get high-resolution optical spectra of the faint outer regions of galaxies. It is the present availability of large optical telescopes, of high-resolution, long-slit spectrographs and of efficient electronic imaging devices that have made our observing program feasible. Six years ago my colleagues and I set out to measure the rotational velocities completely across the luminous disk of suitably tilted spiral galaxies. Our aim was to study the internal dynamics and distribution of mass in individual galaxies as a function of the galaxies' morphology. We have now observed 60 spiral galaxies: 20 each of the three major types Sa, Sb and Sc. We have selected galaxies that have a well-defined type, that are well inclined to the plane of the sky (yielding a large component of orbital velocity along the line of sight), that have an angular diameter no larger than the slit of the spectrograph and that span a large range of luminosities within each type.

Most of the spectra have been obtained with two four-meter telescopes, the one at the Kitt Peak National Observatory in Arizona and the one at the Cerro Tololo Inter-American Observatory in Chile. A few of the spectra were recorded with the 2.5-meter telescope at the Las Campanas Observatory in Chile.

After the photons from the galactic source pass through the slit of the spectrograph and are dispersed by a diffraction grating, they are focused on a "Carnegie" image tube (RCA C33063), where they are multiplied by a factor of 10 or more before they are recorded by the photographic emulsion. Exposures of two to three hours are recorded on Kodak IIIa-J plates whose sensitivity, matched to that of the image tube, has been much increased by having previously been baked at 65 degrees Celsius for two hours in a special "forming" gas (nitrogen with an admixture of 2 percent hydrogen) and preexposed to flashes of light. Without the image tube and

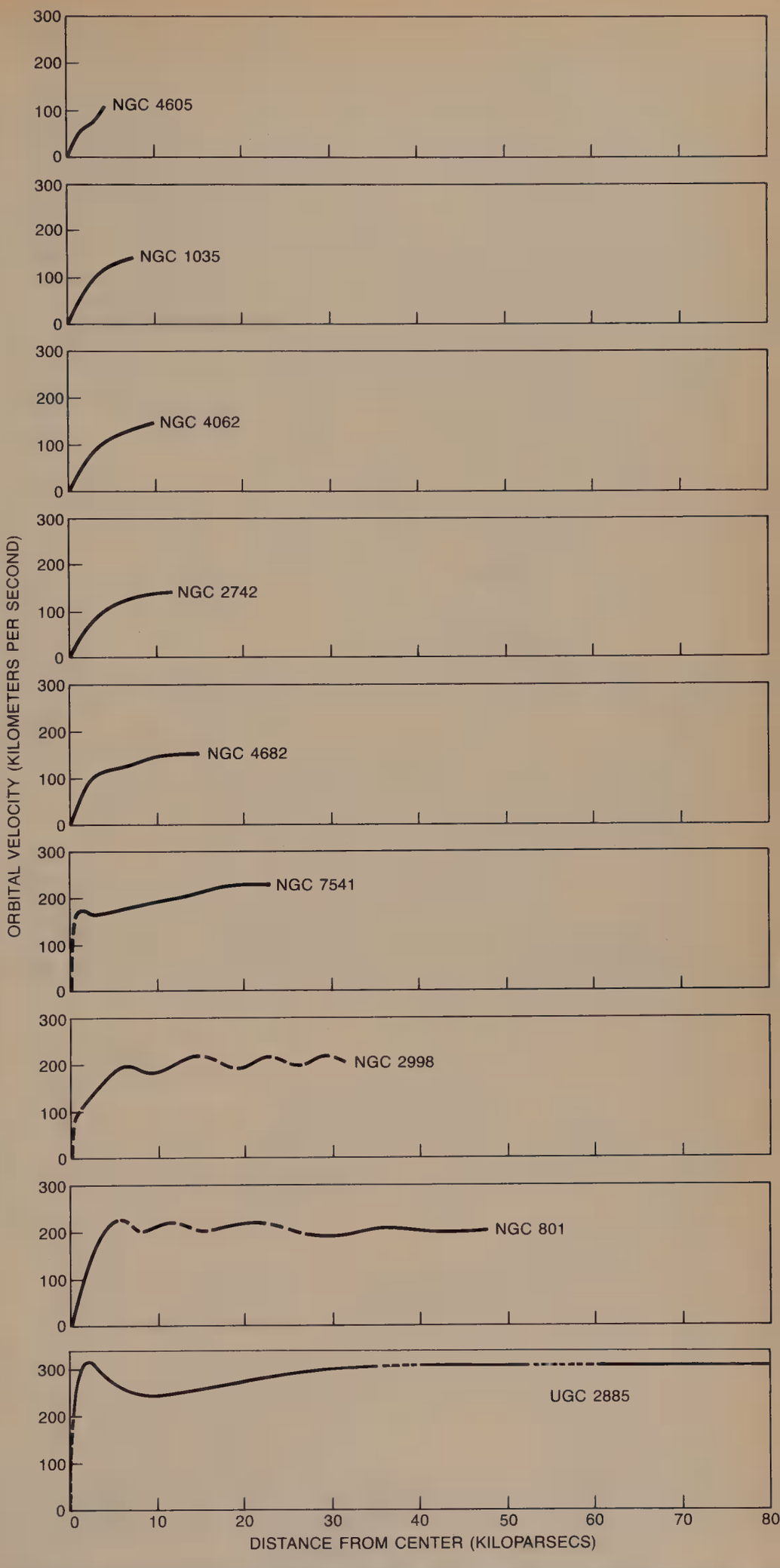
the plate-sensitizing methods exposure times would have been prohibitively long: from 20 to 60 hours.

Generally two exposures are made of each galaxy. In one exposure the spectrograph slit is made to coincide with the major (long) axis of the galaxy; each point on the spectrum arises from a single region of the galactic disk. The Doppler, or velocity, displacements of the emission lines are readily discerned in the developed image. A second exposure is made with the spectrograph slit aligned with the minor axis of the galactic disk. Since the orbital velocities are now perpendicular to the line of sight, no Doppler shifts are evident. The absence of line displacements with the slit along the minor axis is confirming evidence that the motions we study are indeed orbital ones.

In order to have a reference scale against which to measure the displacement of emission lines in galactic spectra astronomers formerly recorded neon lines from a lamp along the edges of the spectrum. We have now dispensed with this procedure. Instead we measure displacements directly from the unshifted lines on each plate that are emitted by hydroxyl (OH) molecules in the earth's atmosphere. Many astronomers have adopted sophisticated plate-scanning devices to measure line positions, particularly for faint signals. We, however, still measure the location of the emission lines with the aid of a microscope whose stage can be moved in two directions. We are able to measure positions in each coordinate to the high accuracy of one micrometer.

In our work we define the nominal radius of a galaxy as that distance at which the surface brightness of the galaxy has fallen to the threshold of detectability on plates made with the 48-inch Schmidt telescope on Palomar Mountain, a value equal to 25th magnitude per square second of arc. For establishing the distance to the objects examined, and hence their actual size, we adopt a value for the Hubble constant (which specifies the expansion rate of the universe) of 50 kilometers per second per megaparsec. (A megaparsec is 3.26 million light-years.)

From the measured velocities of the strongest emission lines we compute a smooth rotation curve by averaging together the approaching and receding velocities from the two sides of the galactic disk. Although each galaxy exhibits distinctive features in its rotational pattern, the systematic trends that emerge are impressive. With increasing luminosity galaxies are bigger, orbital velocities are higher and the velocity gradient across the nuclear bulge is steeper. Moreover, each type of galaxy displays characteristic rotational prop-



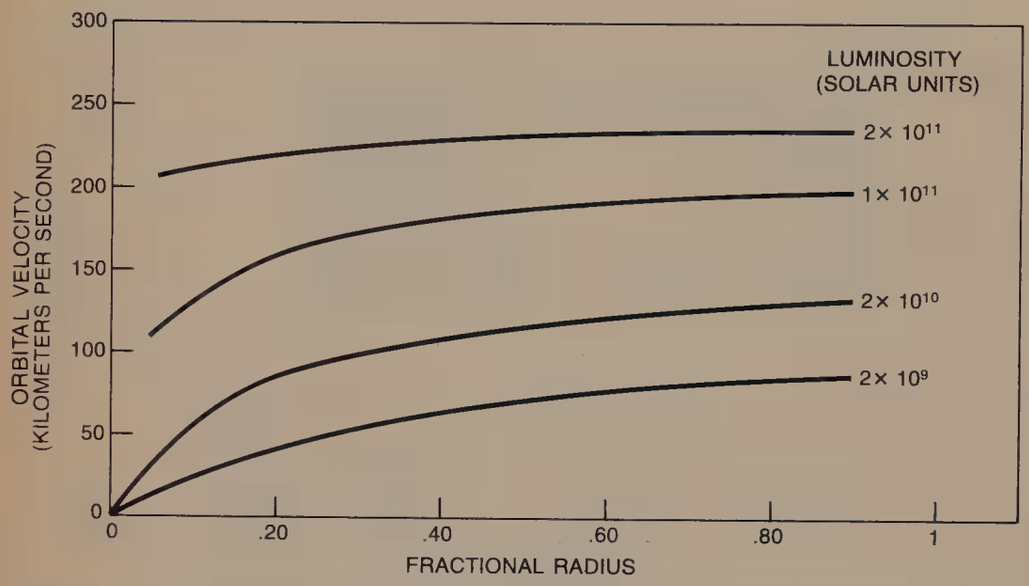
ROTATION CURVES show orbital velocities of nine Sc galaxies from the center outward. Galaxies increase in luminosity from top to bottom. With increasing luminosity galaxies are larger, orbital velocities are higher and velocity gradients near the galactic center are steeper.

erties. For example, the most luminous Sa galaxies rotate more than 50 percent faster at the midpoint of their radius than equally luminous Sc galaxies. Among Sc galaxies the most luminous rotate more than twice as fast at comparable radial distances as Sc galaxies that are only a hundredth as luminous.

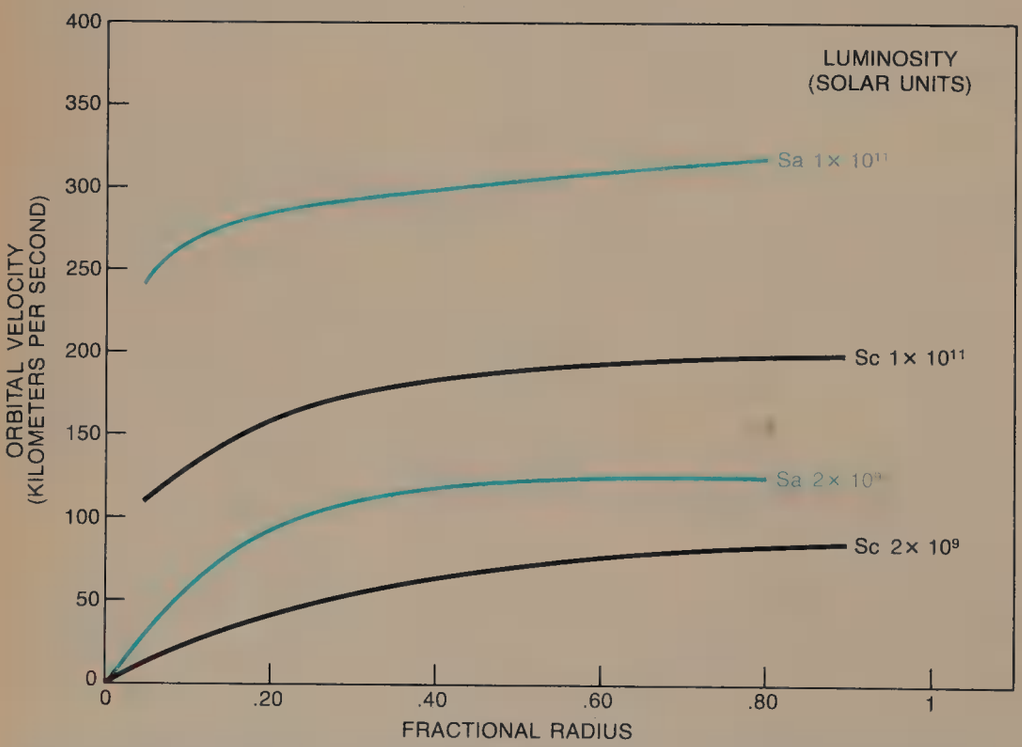
One overwhelming conclusion emerges from our observations. Virtually all the rotation curves are either flat or rising out to the visible limits of the galaxy. There are no extensive regions where the velocities fall off with distance from the center, as would be predicted if mass were centrally concentrated. The conclusion is inescapable: mass, unlike luminosity, is not concentrated near the center of spiral galaxies. Thus the light distribution in a galaxy is not at all a guide to mass distribution.

On the basis of their rotational velocities the masses of the galaxies in our study range from 6×10^9 to 2×10^{12} times the mass of the sun inside their optical radius. We cannot yet specify the total mass of any one galaxy because we do not see any "edge" to the mass. Instead the mass inside any given radial distance is increasing linearly with distance and, contrary to what one might expect, is not converging to a limiting mass at the edge of the visible disk. The linear increase of mass with radius indicates that each successive shell of matter in the galaxy must contain just as much mass as every other shell of the same thickness. Since the volume of each successive shell increases as the square of the radius, the density of matter in successive shells must decrease as 1 over the radius squared in order for the product of the density times the volume to remain constant.

The theoretical model that least disturbs generally accepted ideas about galaxies accounts for the observed rotation curves by embedding each spiral galaxy in a spherical "halo" of matter that extends well beyond the visible limits of the galactic disk. The gravitational attraction of this unseen mass keeps the orbital velocities of the galaxies from decreasing with distance from the galactic center. It is perhaps disappointing that the observations yield almost no information on the detailed distribution of the invisible dark matter. One can nonetheless say that the dark matter is not part of the overall background density of matter in the universe but rather is strongly clumped around galaxies. This is evident because the density of nonluminous matter decreases, albeit slowly, with distance from the galactic center, and the density even at large radial distances is between 100 and 1,000 times higher than the mean density of the universe.



ORBITAL VELOCITIES are depicted schematically for Sc galaxies of varying luminosity as a function of the optically visible radius of each galaxy. Luminosities in solar units differ by two orders of magnitude. At every radial distance orbital velocities increase with luminosity.



COMPARISON OF Sa AND Sc GALAXIES shows that for equal luminosity orbital velocities are significantly higher in Sa galaxies than they are in Sc galaxies at every radial distance. This implies that Sa galaxies harbor more mass per unit of luminosity than Sc galaxies do.

Although there are other models that try to account for the high orbital velocities, all are less satisfactory than a single halo of dark matter. If all the required unseen matter is put in a disk, the disk will quickly become unstable and form itself into a bar. The important finding that halos are necessary for stabilizing a disk was first elucidated by Jeremiah P. Ostriker and P. J. E. Peebles of Princeton University.

The observed dynamic effects are reproduced by models of spiral galaxies that put the mass in a nucleus, a surrounding bulge, a disk and a halo. Particularly interesting models have been developed by John N. Bahcall and Raymond M. Soneira of the Institute for Advanced Study, Maarten Schmidt of Cal Tech and S. Casertano of the Scuola Normale Superiore in Pisa. Perhaps the most radical idea for explaining the observed high rotational velocities is one advanced independently by Joel E. Tohline of Louisiana State University and M. Milgrom and J. Bekenstein of the Weizmann Institute of Science. They have proposed that at great distances the Newtonian theory of gravitation must be modified, thereby allowing rotational velocities in galaxies to remain high at such distances from the galactic center even in the absence of unseen mass.

Additional evidence on the high rotational velocities of matter in spiral galaxies is provided by the 21-centimeter radio waves emitted by the neutral (unionized) hydrogen in the galactic disk.

Early studies of the 21-centimeter radiation of a few spiral galaxies by Morton S. Roberts of the National Radio Astronomy Observatory showed that the rotational velocities of the hydrogen are high. With multiple radio telescopes, notably the array at Westerbork in the Netherlands and the Very Large Array at Socorro, N.M., it is possible to match and even exceed the resolving power of optical telescopes and thereby to study the distribution of hydrogen in galaxies similar to those we have observed. Albert Bosma of the State University of Leiden has shown for a wide variety of galaxy types that the orbital velocities of neutral hydrogen remain high at large distances from the galactic center.

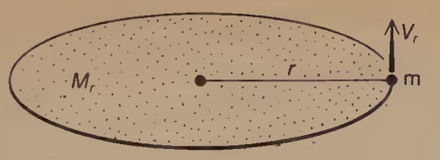
In general the apparent diameters of galaxies are similar whether they are measured by optical observations or by radio ones. For a small set of galaxies, however, hydrogen extends several times as far out from the center as the luminous stars do. For such objects it is possible to determine the gravitational potential beyond the limits of the optically visible galaxy. In several instances the hydrogen does not remain in a plane but is warped sharply near the edge of the visible disk. It is therefore not certain whether the gas velocities that have been measured at the largest distances from the center are true circular orbital velocities or whether they represent more complex motions.

Renzo Sancisi of the University of Groningen, who has studied such warped galaxies, has suggested that the orbital velocities may in fact be decreasing beyond the limits of the visible galaxy. The velocities, however, seem to decrease only slightly, perhaps by 20 kilometers per second, or about 10 percent, and then hold constant at that value at larger distances. The radio observations are continuing and should offer important information on the far outer regions of galaxies.

Students of galaxies are fortunate in being able to examine the properties of galaxies a long way off and then return to the galaxy where they live and ask if it exhibits the same features as other galaxies. It was not so long ago that astronomers believed the sun, about eight kiloparsecs from the center of our galaxy, was near the edge of it and that the galaxy was only of moderate size. Now all the evidence indicates that our galaxy too extends well beyond the position of the sun and that its mass continues to increase.

The velocity of the sun in its orbit around the center of the galaxy is placed at 220 kilometers per second by James E. Gunn and Gillian R. Knapp of Princeton and Scott D. Tremaine of the Massachusetts Institute of Technology. Other estimates run as high as 260 kilometers per second. At the lower value the amount of mass between the sun and the center of the galaxy is about 10^{11} solar masses. On the evidence that substantial mass lies beyond the sun's distance from the galactic center, the galactic mass out to 100 kiloparsecs may reach 10^{12} solar masses, which would place our galaxy in a class with the largest galaxies of its type.

Some 30 years ago Jan H. Oort of the Leiden Observatory demonstrated that the observable mass of stars and gas in the galactic disk in the vicinity of the sun is too low by almost a factor of two to

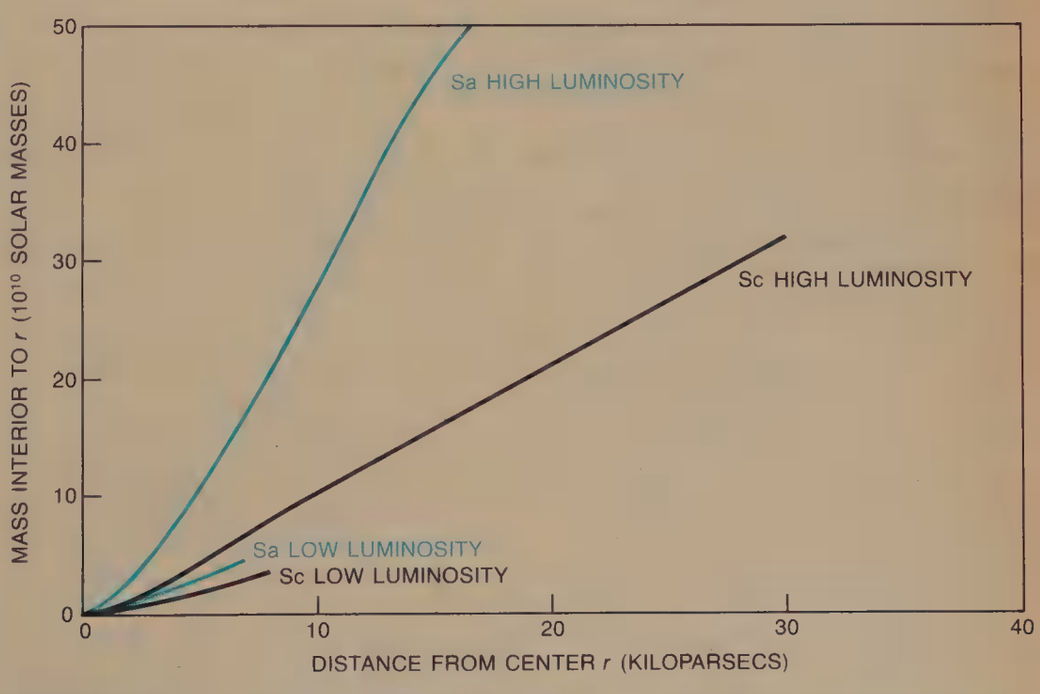


$$\frac{GmM_r}{r^2} = \frac{mV_r^2}{r} \rightarrow M_r = \frac{rV_r^2}{G}$$

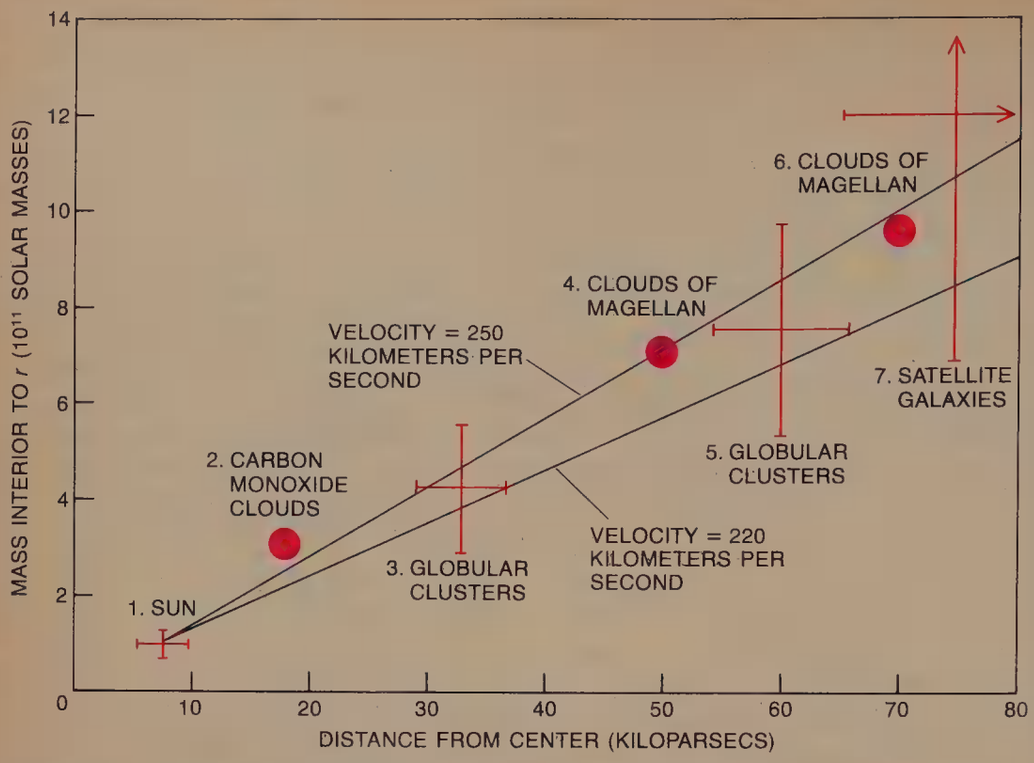
GRAVITATIONAL FORCE = CENTRIFUGAL FORCE

NGC 1035			NGC 2998	
RADIUS (KILOPARSECS)	VELOCITY V_r (KILOMETERS PER SECOND)	INTERIOR MASS M_r (10^{10} SOLAR MASSES)	VELOCITY V_r (KILOMETERS PER SECOND)	INTERIOR MASS M_r (10^{10} SOLAR MASSES)
.5	39	.018	87	.088
1	65	.098	102	.24
2	91	.39	126	.74
3	107	.80	142	1.4
5	123	1.8	182	3.9
8	135	3.4	204	7.7
20			214	21
30			214	32

MASS INSIDE A GIVEN RADIAL DISTANCE can be calculated from the equivalence of gravitational force and centrifugal force at distance r from the center of the galaxy. In the equations G is the constant of gravitation, m is the mass at distance r, M_r is the mass inside r and V_r is the orbital velocity of mass m. The mass inside r increases linearly with distance. The table gives the mass inside r for two Sc galaxies: NGC 1035, of low luminosity, and NGC 2998, of high luminosity. At every distance from the galactic center the more luminous galaxy exhibits higher orbital velocity and therefore must have much more mass inside that distance.



COMPARISONS OF INTERIOR MASS for both low- and high-luminosity galaxies show that the mass rises with approximate linearity with distance r from the center and gives no sign of approaching a limit at the edge of the optically visible galaxy. At every radial distance Sa galaxies exhibit higher mass and therefore higher density than Sc galaxies of equal luminosity.



STUDIES OF OUR OWN GALAXY yield estimates of the mass inside *r* measured in kiloparsecs from the galactic center. The estimates are calculated from the orbital velocities and mean distances of a variety of objects. The value adopted for the orbital velocity of the sun at a distance of eight kiloparsecs is 220 kilometers per second. The second point is obtained from the mean velocity of carbon monoxide clouds at a mean distance of 18 kiloparsecs, measured by Leo Blitz of the University of Maryland at College Park. The third and fifth points are derived from the velocity of globular clusters of stars in the "halo" of our galaxy at two different average distances from the nucleus. The velocities of the nearer clusters were analyzed by Carlos Frenk, who was then working at the University of Cambridge, and Simon White of the University of California at Berkeley. The velocities of the more distant clusters were studied by F. D. A. Hartwick of the University of Victoria and Wallace L. W. Sargent of the California Institute of Technology. The fourth point was obtained from velocities of the Clouds of Magellan, the nearest galaxies to our own, as estimated by Tadayuki Murai and Mitsuaki Fujimoto of Nagoya University. The sixth point represents independent estimates of the distance and velocities of the Clouds of Magellan made by D. N. C. Lin of the Lick Observatory and Donald Lynden-Bell of the University of Cambridge. The final point is based on the velocity of more remote satellite galaxies as estimated by Jaan Einasto and his colleagues at the Estonian S.S.R. Academy of Sciences. The extent of the vertical lines indicates the range of values for orbits of different geometries. The measurements suggest that rotational velocities in our galaxy lie between 220 and 250 kilometers per second and remain constant out to 80 kiloparsecs, or roughly 10 times the sun's distance from the galactic center. The mass inside 80 kiloparsecs is likewise some 10 times the mass inside the radial distance of the sun, or about 10^{12} solar masses.

account for the disk's gravitational attraction on the stars far out of its central plane. This study offered the first evidence that our galaxy too harbors mass that is not luminous.

More recent evidence comes from the orbital velocities of objects in the plane of the galaxy considerably farther out than the sun. Measurements are difficult, but the velocities have been deduced for a few special cases. For example, Leo Blitz of the University of Maryland at College Park has determined the velocities of clouds of carbon monoxide at a distance of nearly 16 kiloparsecs from the galactic center. These velocities, together with the velocities of hydrogen clouds determined by Blitz and Shrinivas Kulkarni and Carl E. Heiles of the University of California at Berkeley, yield a rotation curve that continues to rise with increasing distance from the galactic center.

In order to deduce the mass at still larger distances the velocities of globular star clusters in the halo of our galaxy, with one sample of clusters at 30 kiloparsecs from the center and another at 60 kiloparsecs, have been measured by F. D. A. Hartwick of the University of Victoria, Wallace L. W. Sargent of Cal Tech, Carlos Frenk of the University of Cambridge and Simon White of the University of California at Berkeley. Their work shows that the mass continues to increase with approximate linearity to the mean distance of the clusters.

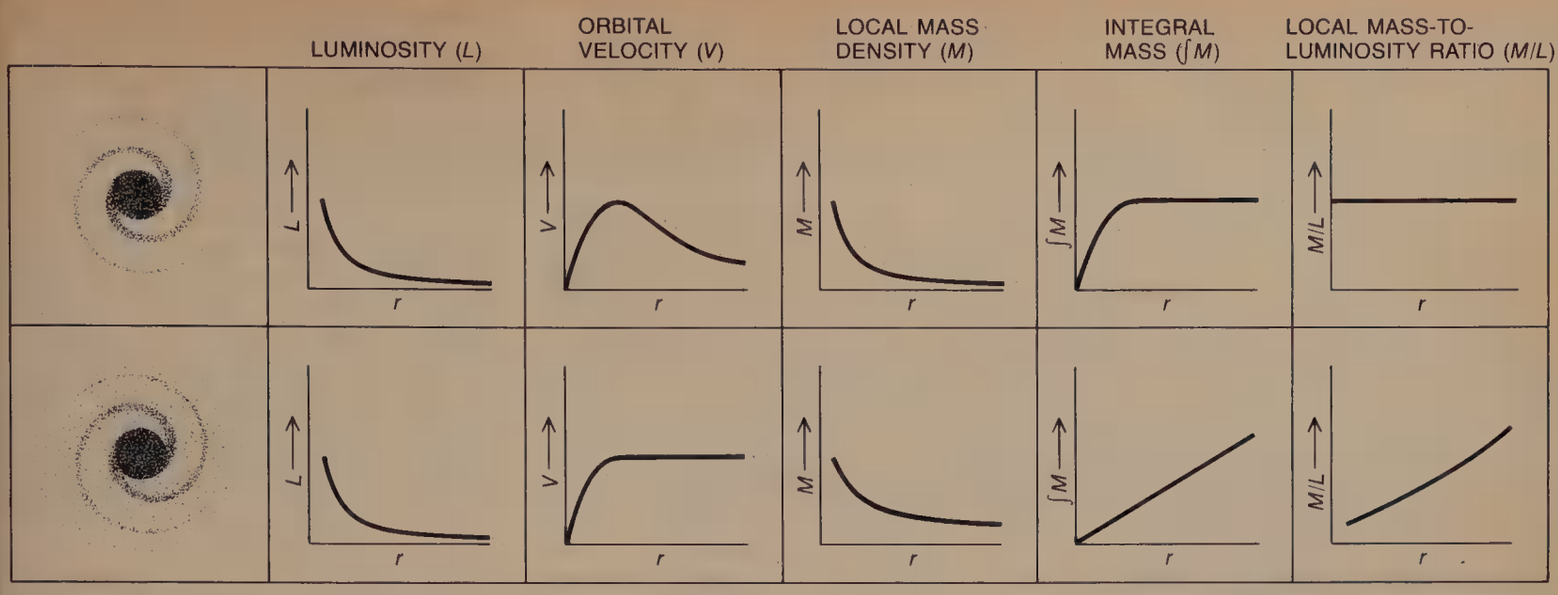
With effort and imagination it is possible to sample the gravitational potential at even more remote distances. Our galaxy is not alone in intergalactic space; it has a retinue of smaller satellite galaxies. The orbits of the two closest satellites, the Large and Small Clouds of Magellan, a little less than 60 kiloparsecs from the center of our galaxy, are highly uncertain. Model orbits have been proposed, however, by Tadayuki Murai and Mitsuaki Fujimoto of Nagoya University, D. N. C. Lin of the Lick Observatory and Donald Lynden-Bell of the University of Cambridge. From the model orbits they deduce values of mass that are consistent with those yielded by the globular clusters.

For still greater distances Jaan Einasto and his colleagues at the Estonian S.S.R. Academy of Sciences have relied on a combination of enormously distant globular clusters and satellite galaxies to deduce the mass to distances beyond 80 kiloparsecs. When the results from such analyses are combined, they indicate a galaxy in which orbital velocities remain in the range of 220 to 250 kilometers per second out to almost 10 times the distance of the sun from the galactic center. Such a mass distribution is mandatory if our galaxy is to resemble all the other spiral galaxies my colleagues and I have studied. It moves the sun from a relatively rural position to a much more urban one.

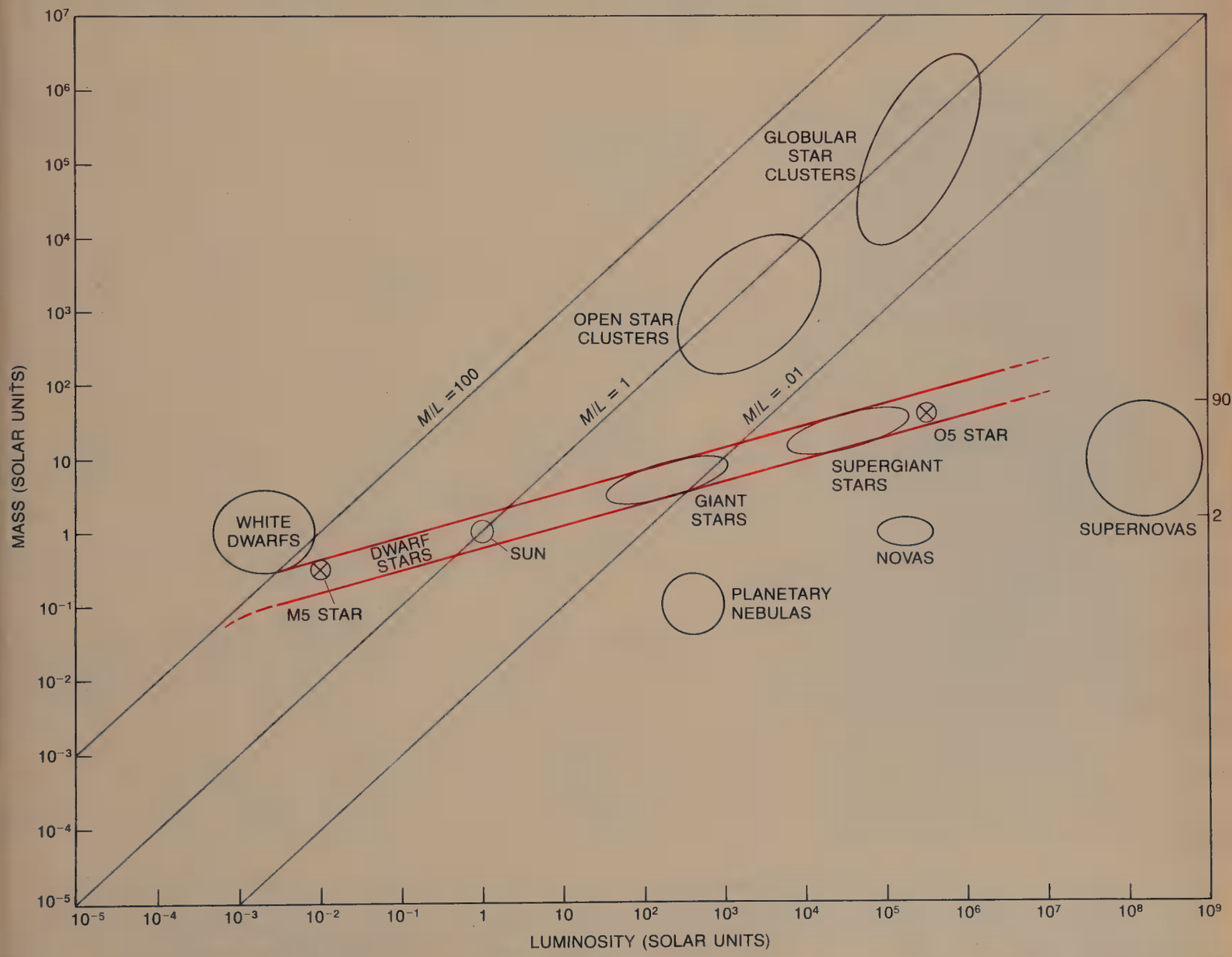
The broad conclusion that can be drawn from all these results is that as the disk of a spiral galaxy is scanned from the center outward the total mass of luminous and dark material falls off slowly and the luminosity (measured in the blue region of the spectrum) falls off rapidly. As a result the ratio of the local mass density to the local (blue) luminosity density, which can be expressed for convenience as the value of the ratio M/L, increases steadily with distance from the galactic center. In the nuclear region a lot of luminosity is produced by relatively little mass, whereas at large distances little luminosity is produced by a lot of mass. If there were no invisible material clumped around galaxies, the mass distribution would simply follow the luminosity distribution and the M/L ratio would be approximately constant across the disk from its center to its edge.

If mass and luminosity are measured in units of solar mass and solar luminosity, the M/L ratio of the sun is of course 1/1. In such units (omitting the denominator, which is simply 1) the average M/L ratio near the nucleus of a spiral galaxy has sunlike values of 1 or perhaps 2 or 3. Toward the edge of the visible disk, as luminosity decreases, the M/L value climbs to 10 or 20. Beyond the visible disk, where the luminosity falls essentially to zero and the mass remains high, the average M/L value soars into the hundreds.

In an effort to identify the constituents of the invisible halo we must ask what celestial objects have high values of M/L. Stars like the sun are clearly ruled out. The luminous hot young stars that delineate a galaxy's spiral arms are even poorer candidates; their M/L values are about 10^{-4}. At the other extreme the old red-dwarf stars that populate the nuclear bulge and the outlying regions of the galaxy have both a low mass and a



HYPOTHETICAL AND ACTUAL GALAXIES deviate sharply in all their properties except luminosity. The typical actual spiral galaxy at the bottom has a massive nonluminous halo. The hypothetical galaxy at the top has no halo. Its surface brightness decreases rapidly, orbital velocities outside the nucleus decrease in Keplerian fashion, local mass density falls in parallel with luminosity, integral mass reaches a limiting value and the ratio of mass to luminosity stays approximately constant with increasing radial distance. Such were the expected properties of a galaxy. In an actual galaxy the presence of a dark halo changes everything but the galaxy's optical appearance. The orbital velocities remain high, the local mass density falls only slowly, the integral mass increases linearly with radius and the mass-to-luminosity ratio steadily increases as the halo of the galaxy contributes more mass and the luminous disk falls to the threshold of detectability.



MASS AND LUMINOSITY ARE COMPARED for typical components of a spiral galaxy such as our own. The mass and luminosity of the sun are taken as unity. In solar units the value of the ratio of mass to luminosity, M/L, for normal stars decreases from about 30 for cool, old dwarf stars (type M5) to about 10^{-4} for hot, young stars (type O5). Only extremely dense white-dwarf stars have an M/L value in excess of 100. Some other class of objects is needed to populate the halo of a galaxy, where M/L values soar into the hundreds.

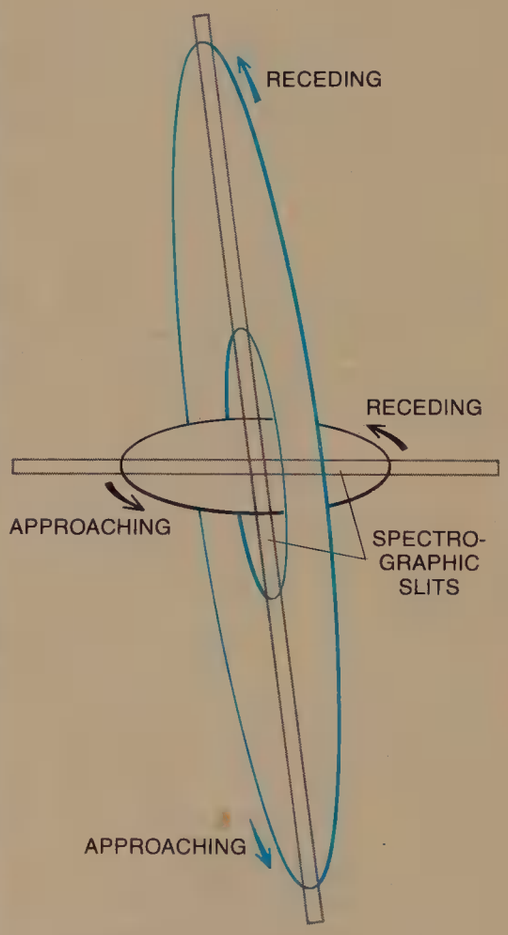
low blue luminosity. Their M/L values, about 20, are still far short of the values needed for the halo. Moreover, a halo consisting of very-low-mass red dwarfs would reveal its presence by radiating strongly in the infrared region of the spectrum. All attempts to detect a halo by its visual, infrared, radio or X-ray radiation have failed.

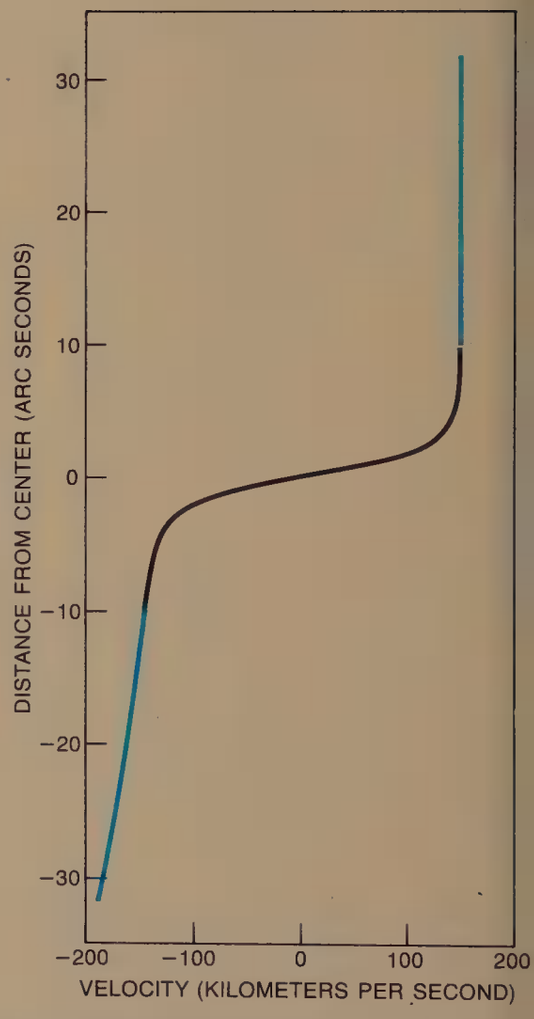
What candidates are left? Normal stars radiate energy generated by thermonuclear processes, which convert hydrogen and helium into heavier elements. Such nuclear processes are kindled only in bodies whose mass is large enough for the gravitational energy to raise the temperature at the core of the star to several million degrees Kelvin (degrees C. above absolute zero). The minimum mass required is about .085 times the mass of the sun. Jupiter, the largest planet in the solar system, falls short of this value by a factor of nearly 100. A halo of planetlike bodies, perhaps protostars that failed to become stars, is at least conceivable, although rather unlikely. In sum, the only requirement for the halo is the presence of matter in any cold, dark form that meets the M/L constraint, from neutrinos to black holes.

So far I have described the rotational properties of relatively isolated normal spiral galaxies. There is additional observational evidence for high M/L ratios at large distances from the nuclei of other galaxies. Occasionally nature offers an unexpected opportunity to probe its secrets. Recently François Schweizer of the Carnegie Institution, Bradley C. Whitmore of Arizona State University and I have been fascinated by the faint "anonymous" galaxy AO 136–0801, one of a class of spindle galaxies with polar rings. It is called anonymous because it is not listed in any of the standard galactic catalogues; its numerical designation corresponds to its location in the sky.

Our observations of the distribution of light across the spindle show that it is a low-luminosity disk of stars viewed nearly edge on, with little or no gas and dust and no spiral structure. Such galaxies are classed as SO galaxies and represent a significant fraction of all disk galaxies. By our usual methods we have determined the rotational properties of the disk by measuring the Doppler shift of absorption lines from its component stars. A short distance from the center of the object along the major axis of the spindle rotational velocities reach 145 kilometers per second, a value that corresponds closely to velocities measured in type-Sa galaxies of low luminosity. Along the minor axis the orbital velocities show no line-of-sight component, confirming evidence that we are observing a rotating disk of stars.

The unusual feature of AO 136–0801 is a large ring, also seen nearly edge on, that encircles the narrow axis of the spindle by passing almost over the disk's center of rotation [*see illustration below*]. The ring is composed of gas, dust and luminous young stars. The gas reveals itself by its emission-line spectrum, the dust by its absorbing effects where it crosses in front of the spindle





DISK INSIDE A RING is exhibited by the "anonymous" galaxy AO 136–0801, so called because it is not listed in standard catalogues. (The numbers give its position in right ascension and declination.) The oval central region is a rotating disk of stars seen nearly edge on. The stars and gas in the thin ring are also rotating but in a plane almost perpendicular to the disk. The directions of rotation are indicated by the diagram in the middle, which shows how the slits of the spectrograph were oriented for measuring orbital velocities in the disk and the ring. The two sets of velocity measurements are plotted at the right. At 10 seconds of arc from the center the velocities in the disk (*black*) and in the ring (*color*) are essentially the same. The velocities in the ring, however, can be measured out to nearly three times the optical radius of the disk, and they remain virtually constant. It appears that the mass is continuing to increase linearly out to distances much greater than the disk radius, and that objects in the ring respond to a gravitational potential that is not disklike but spherical.

SCIENCE/SCOPE

Printed circuit boards made of a new material may permit better direct soldering of large leadless ceramic chip-carriers. A Hughes Aircraft Company study proposes using quartz-fabric-reinforced polyimide resin in place of glass-epoxy or glass-polyimide boards. The new material has nearly the same thermal expansion coefficient as ceramic chip carriers. When a leadless carrier is soldered directly to a quartz-fabric polyimide board, there are no shear stresses caused by heating or cooling. Such stresses often cause solder joints to fail on conventional reinforced printed circuit boards.

A semiautomatic production line makes solar cells that are higher performing and much lower in cost than conventional cells. The new facility at Spectrolab, a Hughes subsidiary, creates highly uniform cells and increases productivity. It uses the latest techniques, including photolithographic masking for contact patterns, with subsequent contact plate-up. The first "mechanized" cells have been delivered to Hughes for the Intelsat VI communications satellites.

Advanced electro-optical devices produced for the military by Hughes today have benefited from a 7-to-1 productivity improvement over the last five years. These devices include laser rangefinders, laser designators, and infrared night vision systems. The rise in productivity stems from more automation, computer-aided manufacturing, and employee involvement in solving day-to-day problems.

A wide-field-of-view head-up display can provide pilots with critical sensor and steering information in low-altitude flights at night and under poor visibility conditions. Head-up displays save a pilot from looking down at his instruments by superimposing such data as airspeed, heading, and target information on a glass-like combiner mounted at the pilot's eye level. Hughes pioneered the technology used in its HUD, which incorporates diffraction optics made through a process involving holographic techniques and lasers. In addition to providing the wide field of view, the display is brighter, more transparent, and doesn't obstruct the pilot's forward vision. The display resists glare, reflections, and hot spots caused by the sun. Another important advantage is its ability to display wide-field-of-view scenes from infrared sensors. Test pilots praised the display during evaluation flights in a U.S. Navy F/A-18 Hornet.

Hughes is seeking engineers to develop advanced systems and components for many different weather and communications satellites, plus the Galileo Jupiter Probe. Immediate openings exist in applications software development, data processing, digital subsystems test, microwave/RF circuit design, power supply design, digital communications, signal processing, spacecraft antenna design, system integration test and evaluation, and TELCO interconnection. Send your resume to Ray Bevacqua, Hughes Space & Communications Group, Dept. SE, Bldg. S/41, M.S. A300, P.O. Box 92919, Los Angeles, CA 90009. Equal opportunity employer.

Creating a new world with electronics

HUGHES

HUGHES AIRCRAFT COMPANY

For more information please write:
P.O. Box 11803, Los Angeles, CA 90291

and the stellar component by its knotty, bluish appearance in photographs. The maximum diameter of the ring is several times greater than the long axis of the spindle. As a consequence the motions of the objects in the ring offer a unique opportunity to probe the gravitational field perpendicular to the galactic disk out to distances exceeding the visible radius of the disk.

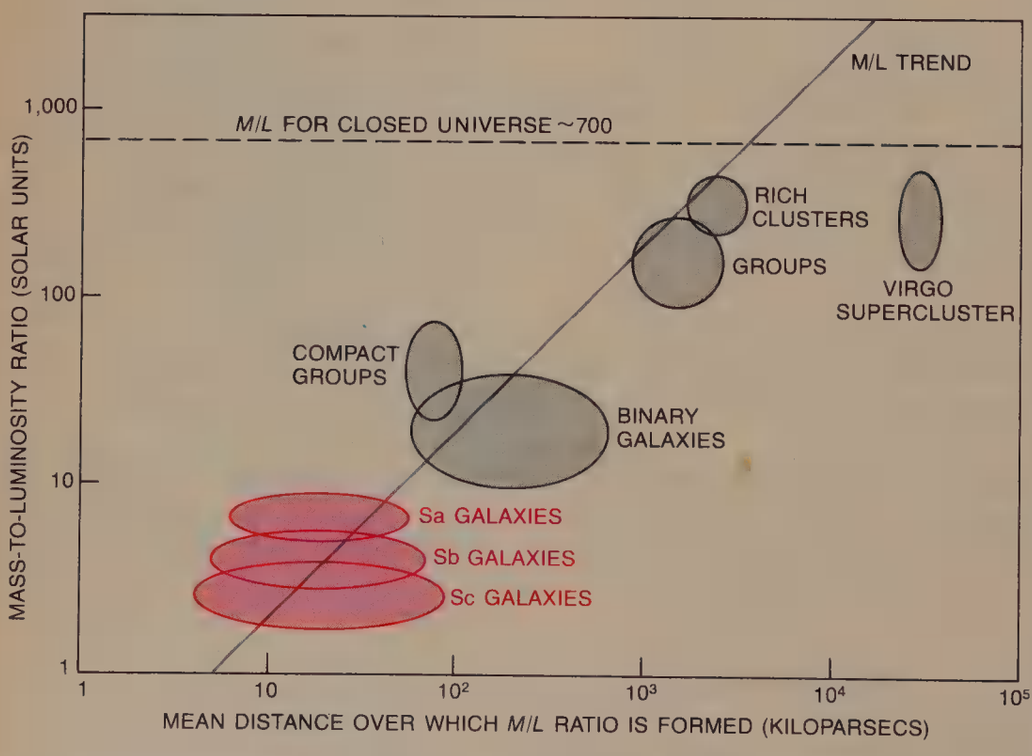
Our spectrographic observations confirm that the ring is indeed rotating at right angles to the plane of rotation of the disk. It seems improbable that this dynamical configuration could have arisen in the normal evolution of an isolated disk galaxy; the configuration must be the result of some kind of event, such as an encounter with another galaxy or with a disk of gas. By measuring the displacement of emission lines we find that the ring's velocity of rotation is about 170 kilometers per second and that the velocity curve is flat or slightly rising out to a distance of almost three times the radius of the inner disk. If the velocity curves of the disk and the ring are plotted on the same velocity-distance scale, the two are seen to have nearly identical values at the same distance from the center of the galaxy. The high rotational velocity of the ring offers strong evidence for the existence of a massive halo extending at least three times farther than the visible radius of the disk. Moreover, the shape of the halo must be more nearly spherical than disklike. Calculations show that if the halo were as flat as the disk, the velocities above the plane of the disk would be smaller than those in the disk by 20 to 40 percent.

I have been describing determinations of mass made by measuring the velocity of orbiting test objects, objects in the central disk of a galaxy and objects orbiting the pole of an unusual galaxy. Other special instances can help to shed light on the quantity of dark matter in the universe. Galaxies often exist in pairs. In such instances one galaxy can be considered a test object in orbit around the other. The analysis of such a system is complex because both the orientation of the orbit in space and the position of the galaxy in the orbit are unknown. One can, however, resort to the observed properties in a large sample of double galaxies (the difference between the velocities of the two galaxies, their angular separation and their luminosity) to infer from statistical arguments the probable distribution of orbital elements and M/L ratios appropriate to the galaxies.

Independent analyses by Edwin L. Turner of Princeton, Steven D. Peterson, working at Cornell University, Linda Y. Schweizer of the Carnegie Institution and I. D. Karachentsev of the Special Astrophysical Observatory in the U.S.S.R. yield mean M/L values in the range between 25 and 100. These values of M/L are an average over a distance equal to the separation of the galaxies in each pair, a distance generally equal to several galaxy diameters, or on the order of 100 kiloparsecs. This result helps to confirm the view that halos of dark matter, with large values of M/L, extend well beyond the optical limits of galaxies.

We can now return to our original question: Does the universe contain enough invisible matter to raise the average density to 5×10^{-30} gram per cubic centimeter, the value needed to close the universe and bring its expansion to a halt? As we have seen, such a density would be reached if the density of nonluminous matter exceeded the density of luminous matter by a factor of about 70. Alternatively what would be needed to close the universe can be expressed in terms of the ratio of total mass to luminosity. That value is roughly 700, compared with 1 for the sun.

Is there any evidence that the M/L value of 700 is approached? Averaged over the visible disks of spiral galaxies, the ratio of total mass (luminous and nonluminous) to luminosity is about 5. For SO and elliptical galaxies the M/L value is higher, on the order of 10. For double galaxies and small groups of galaxies the M/L value increases to between 50 and 100. Analyses of galaxy motions in large clusters indicate M/L values of several hundreds. This increase in mean M/L value with increasing distance from the center of the system was first stressed a decade ago by Einasto, Ants Kaasik and Enn Saar of the Estonian S.S.R. Academy of Sciences, and also by Ostriker and Peebles and by Amos Yahil of the State University of New York at Stony Brook. So far there is no evidence for the existence of M/L values above the critical one of 700 needed to close the universe. The highest of the derived values, however, comes tantalizingly close. Some physicists consider it significant that the inferred values seem to be converging on the critical one rather than being orders of magnitude either higher or lower.



RATIOS OF MASS TO LUMINOSITY are plotted for aggregates of matter on progressively larger scales. The plot is based on one devised by Herbert J. Rood of the Institute for Advanced Study. The M/L value of a density of matter sufficient to arrest the expansion of the universe is about 700. For galaxies values are below 10. The value rises with the size of the aggregate.

Investigations encompassing gigantic distances and vast time scales are made more difficult by this new realization that the distribution of light is an unreliable guide to the distribution of mass in the universe. An unknown fraction of the mass in a spiral galaxy is hidden in a nonluminous constituent, and so is an unknown fraction of the mass in clusters of galaxies. One cannot yet state whether regions of the universe that are devoid of galaxies are simply voids of light or are voids of mass as well. To answer this question astronomers will have to be clever in devising novel observing techniques and physicists will have to determine the properties of exotic forms of matter. Only then will it be possible to establish the nature of the ubiquitous dark matter, to determine the full dimensions and mass of galaxies and to assay the likely fate of the universe.

Oceanus®

The International Magazine of Marine Science and Policy

With offices in one of the world's leading marine research centers and published by the

Woods Hole
Oceanographic Institution,

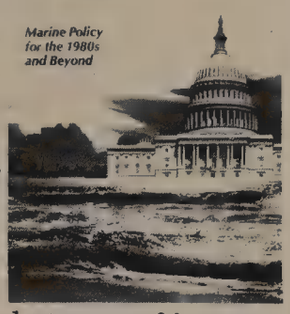
Oceanus magazine is in a uniquely favorable position to monitor significant ocean research and related policy issues.

A case in point is the recent discoveries of deep ocean hydrothermal vents at various locations in the Pacific Ocean. These discoveries are causing a revolution in Oceanography.

Many long-standing biological and geological theories are being revised as a result of this exciting research. Exploration of the vents by the submersible Alvin is turning up new life forms (derived chemosynthetically rather than photosynthetically) and possible new sources of metals, such as zinc, iron, and copper. Scientists are getting a first-hand look at how the earth breathes while garnering supporting evidence for the theories of sea-floor spreading and plate tectonics.

Economically, it is a time of great challenge, too. The cost of running submersibles and support ships, for example, is high and rising. And federal funding for marine research all the while is shrinking. Many critical decisions must be made in nearly every area of oceanic concern.

Oceanus is meeting this challenge by publishing articles that help our readers grasp the significance of present research and that expose them to the substance of important public policy questions.



Recent thematic issues have dealt with:

- Deep Sea Mining
- Research Vessels
- Sharks
- Coastal Problems
- Pollution
- Oceanography from Space

Upcoming issues will look at:

- Marine Birds
- Oil & Gas Activities
- Industry and the Sea
- Polar Research

and much, much more.

Handsomely illustrated with many sharp photographs and line art, **Oceanus** is edited for all those with a serious interest in the three-quarters of our planet that is covered by water. The general readership includes many students, educators, representatives from government and industry, environmentalists, conservationists, as well as scientists in the disciplines of marine biology, geology and geophysics, chemistry, ocean engineering, and physical oceanography.

In addition to the thematic editorial material, **Oceanus** also features:

PROFILES of selected oceanographers that give a glimpse of the human side of science.

A CONCERNS SECTION that exposes the reader to controversial subjects and opinions.

A BOOK REVIEW SECTION that gives readers a convenient way to choose new and timely reading material.

LETTERS TO THE EDITOR or a forum for debate on views and thoughts expressed in the magazine.

Won't you join the Editors of **Oceanus** — published quarterly — in the exploration of the last great wilderness on earth, our oceans. Subscribe today!

Oceanus

SUBSCRIPTION ORDER FORM

Please enter my subscription for:

- ☐ one year at $20.00
- ☐ two years at $35.00
- ☐ payment enclosed ☐ bill me

(prepayment preferred)

Please send MY Subscription to:

Name (please print)

Street Address

City State Zip

Please make checks payable to Woods Hole Oceanographic Institution. Checks accompanying foreign orders must be made payable in U.S. currency and drawn on a U.S. bank. (Outside U.S., add $3 per year to domestic rates). Mail to: OCEANUS, Woods Hole Oceanographic Institution, Woods Hole, Massachusetts 02543.

BOTTLENOSE DOLPHIN, one of the small whales trained by the U.S. Navy for underwater tasks, worked for the authors in experiments on the physiology of breath-holding. This dolphin, known to his trainers as Tuffy, submerged on signal and after various intervals responded to a second signal by nudging a buzzer and exhaling into a funnel, so that his remaining oxygen supply could be measured.

The Physiological Ecology of Whales and Porpoises

Like other mammals, these marine species have a high-energy way of life. In pursuing it they evolved particular adaptations of their own, notably the ability to dive deep for long periods

by John W. Kanwisher and Sam H. Ridgway

It is not generally realized that 70 species of entirely marine mammals—the whales and porpoises—play as crucial a role in oceanic ecosystems as the terrestrial mammals do on land. The broad success of the land mammals at the end of the Mesozoic era (some 65 million years ago) was due in large part to their high-energy way of life, including such features as warm-bloodedness and an expanded central nervous system. When some of these terrestrial mammals later filled a marine ecological niche, they evolved further adaptations of the same high-energy type.

How have the whales and porpoises—the Order Cetacea—overcome the many physiological obstacles presented by living in the ocean? Consider, for one thing, their severely stressed life cycle. Initially expelled from a submerged birth canal into water that can be close to the freezing point, the newborn calf must struggle to the surface unaided before it can even take its first vital breath of air. To be suckled by its mother the calf must hold its breath and return underwater. When eventually it can feed itself, it may have to develop an entirely new behavioral repertory: diving in order to find its prey. Finally, once the cetacean is an adult, it must master a further repertory of complex gymnastics before it can reproduce and thereby give rise to the offspring that initiate the next life cycle.

Consider also the surprising diversity of the cetaceans, both in geographic distribution and in size. Whales and porpoises are plentiful in all the oceans, from the Tropics to the edge of the polar ice, both north and south. Many species even migrate seasonally from cold seas to warm ones and back again. The largest adult blue whale is some 20,000 times heavier than the smallest newborn porpoise, yet both large and small species are found in tropical and polar seas. The toothed whales—the odontocetes—include small porpoises such as the harbor porpoise at one end of their size range and the great sperm whale at the other. How is it that the sperm whale is the cetacean that dives the deepest and stays submerged the longest? How does it avoid the "bends" and other physiological problems human beings encounter when they dive to much shallower depths? The porpoises, being smaller than the smallest of their larger toothed-whale cousins, have a much higher surface-to-volume ratio. How do they manage to stay warm in cold waters? Questions of this kind led us to the first of several investigations: the thermal physiology of porpoises.

Our observations of captive porpoises at Marineland in Florida showed us that the animals were able to maintain a stable internal temperature close to the human one and similarly well regulated. Since water removes heat from an object much faster than air does, we found this observation intriguing. We had thought these thermally stressed mammals might exhibit the more variable body temperatures typical of terrestrial hibernators. This not being the case, we were led to calculate the stringency of the thermal problem faced by a small porpoise living in the cold waters of high latitudes.

For the purpose of heat-flow calculations a cetacean may be regarded as a uniformly warm fleshy core surrounded by a surface layer of insulating blubber. The outward flow of heat from the core, encountering the thermal resistance of the blubber layer, must be great enough to maintain the temperature difference between the surrounding water and the warm inner core. It is known that this difference is equal to the product of the thermal resistance of the blubber times the outward flow of metabolic heat generated in the core. When the insulation is insufficient, more body heat must be generated if the animal is to maintain a constant core temperature. For example, if the core temperature of a terrestrial mammal falls by as little as .5 degree Celsius, the mammal will begin to shiver in order to increase its metabolism. Over longer periods any such imbalance can be corrected by increasing the amount of insulation. The land mammal grows a winter coat and the whale in polar waters develops a thicker layer of blubber.

When the animal is active, such an increased thermal barrier may actually be too much of a good thing. For example, when a whale swims fast, the increase in its metabolic activity will cause its core to overheat. The whale then has recourse to a circulatory stratagem. An increase in blood flow near the body surface, particularly through the flippers and flukes, thermally bypasses the insulating blubber and returns the core temperature to normal. In cetaceans generally a steady body temperature is mainly achieved by such changes in blood flow. The same is true of human beings: when they are too warm, more blood is shunted to the surface of the body, producing the flushed appearance that is the opposite of being "blue from the cold."

There are fundamental differences, however, between the physiology of human beings and the physiology of whales. For example, cetaceans have no sweat glands; evaporative cooling in an aquatic environment is impossible. By the same token, for their surface insulation human beings have only a very poor equivalent of blubber. Human divers have now learned to imitate whales by wearing insulation in the form of foam-rubber suits. Ashore human beings overcome the handicap of their essentially tropical origins by covering themselves with, among other things, winter insulation grown by other terrestrial mammals.

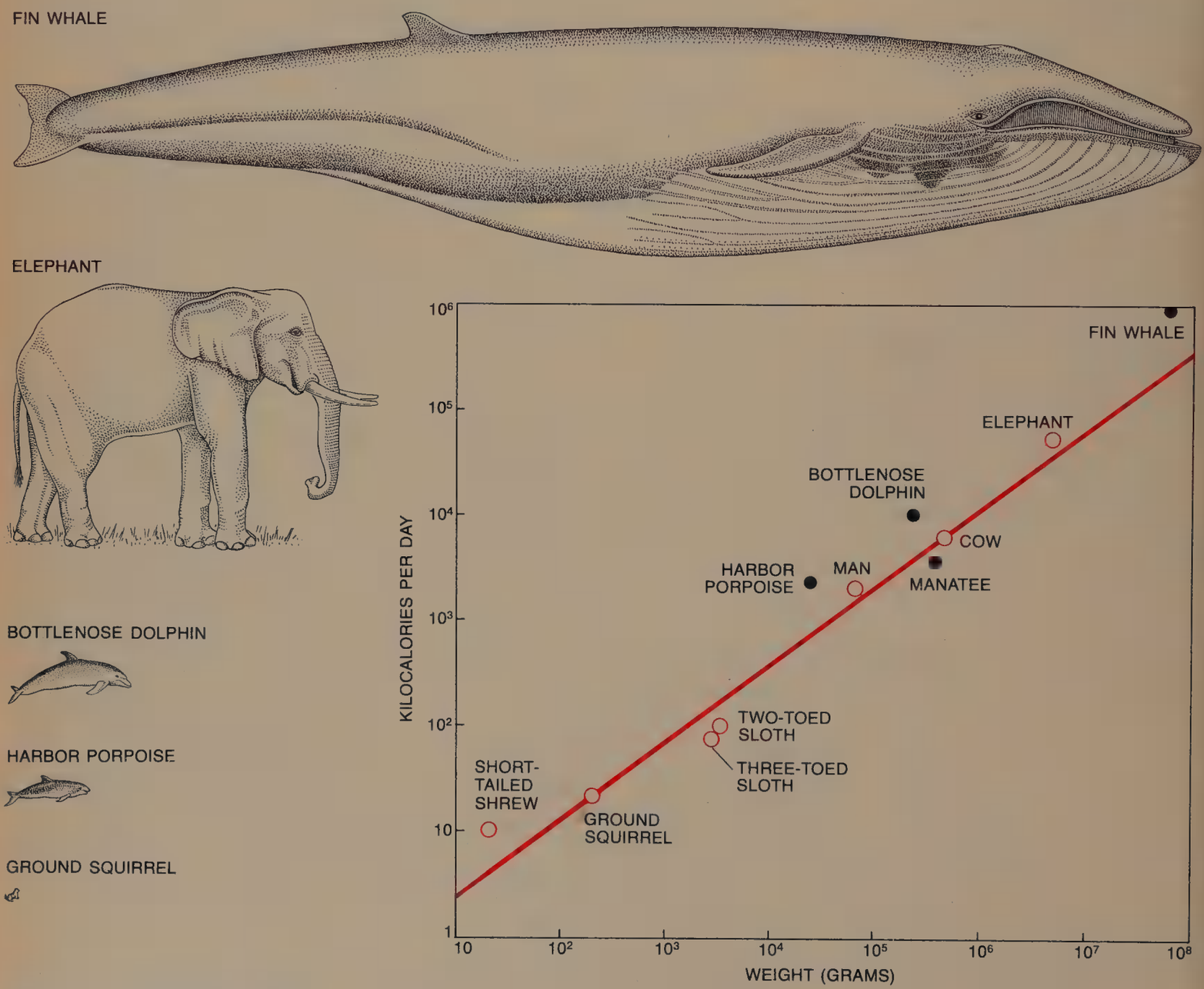
Even before we began our investigations we were well aware that the smallest cetaceans would be the ones that

faced the greatest problem in maintaining a normal mammalian temperature. Considering the total surface area that is constantly losing heat to the surrounding water, the smaller the animal, the smaller the core volume that generates the metabolic heat required to maintain the necessary temperature differential across the blubber. In addition to the disadvantageous geometry of surface-to-volume ratios, considerations of linear scaling also work against any small porpoise trying to stay warm in cold water. Because of their smaller dimensions, the smaller animals will tend to have a thinner layer of blubber.

Regardless of their size, cetaceans maintain a normal mammalian internal temperature of about 37 degrees C. In the Arctic, where the water temperature may be −2 degrees C., the temperature difference across the blubber barrier is nearly 40 degrees. In the Tropics, where the water temperature may be 30 degrees C., the temperature gradient across the barrier would be only a fifth of that in Arctic waters. Thus a porpoise adapted to warm water would seem to be headed for serious thermal trouble on entering polar seas. Yet one of the smallest cetaceans, the harbor porpoise (*Phocaena phocaena*), is found mostly in cold northern waters.

To determine the dimensions of the problem faced by harbor porpoises in cold water we calculated how much metabolic heat they would have to generate in order to keep warm. To determine the total surface area and the thickness of the porpoise's insulating blubber we made direct measurements on a dead harbor porpoise that had washed ashore. When our figures were entered in the formula for heat loss, the only conclusion to be drawn was that with only a normal mammalian metabolic heat supply a porpoise that small could not keep warm in northern waters. Yet there the harbor porpoises were, sporting in the same chill seas that should have killed them.

Such a paradox is upsetting to biologists who have come to expect orderliness when they examine any parameter common to different animal groups. The amount of metabolic heat production in animals of different sizes is one such comparison that seemed well enough understood. For example, the



"MOUSE TO ELEPHANT" CURVE, shown as a solid line (*color*) in the graph at the right, was devised in the 1930's by Francis G. Benedict of the Carnegie Institution of Washington as a rough means of estimating the basal metabolism of various mammals, expressed in kilocalories per day, on the basis of 75 percent of their weight. Large colored circles show how selected terrestrial mammals closely fit the curve, which is constructed on a logarithmic ordinate and abscissa. (The ordinate has a top value of one million kilocalories and the abscissa extends to 100 million grams.) Direct measurement of the metabolism of harbor porpoises (*Phocaena*) and bottlenose dolphins (*Tursiops*) and an estimate for fin whales (*Balaenoptera*) based on lung volume (*black dots*) all lie well above the terrestrial-mammal curve. Weight for weight, however, the metabolism of fin whales is lower than that of such sluggish terrestrial mammals as sloths.

metabolic rates of numerous terrestrial mammals have long been arrayed in what is called the "mouse to elephant" curve. It was believed that given the animal's body weight one could closely estimate from this curve the metabolic heat production of any mammal. Our estimate of the metabolic rate needed to keep the porpoise warm in cold water, however, was so high that it was in clear disagreement with the data from terrestrial mammals.

There were two possible ways to explain the paradox. The first was, quite simply, that the metabolism of the cetaceans does not conform to the terrestrial mammals' mouse-to-elephant curve. This is to say the cetaceans may have an intrinsically higher rate of basal metabolism that allows even a small cetacean to survive in cold water. The second possibility is that cetaceans have a "normal" (which is to say a terrestrial) basal rate of heat production but that by exercising all the time when they are in cold water they can generate enough extra heat to keep themselves from getting too cold. To us the only way to choose between these two explanations was to present the question to the porpoises. Which of these thermal strategies had evolution chosen for them?

With the skilled help of Gunnar Sundnes, a physiologist at the Havforskningsinstitutt in Bergen, we set up a heat-loss experiment. First we learned how to keep young harbor porpoises more or less contented in captivity in a large tank. We then transferred them for testing to a smaller box, filled with seawater, that served as a whole-body calorimeter. The young of the species weigh as little as 15 kilograms. Although they are born with blubber that is only one centimeter thick, the difference between their normal body temperature and the temperature of the surrounding seawater can be as much as 25 degrees C.

The porpoise being tested would lie quietly in its box as the metabolic heat it was generating passed through its insulating layer of blubber and warmed the surrounding water. We found we could put a thermistor in the box and with a little amplification produce a curve plotting temperature against time on a chart recorder that was sensitive to within .001 degree C. Taking into account the volume of water in the box, the slope of the temperature curve provided a running record of how much metabolic heat the porpoise was losing at any moment.

Our results proved to be quite clearly in favor of the first of our two possible explanations: these small porpoises had a metabolic rate two to three times higher than the rate of terrestrial mammals of the same weight. This increased heat production is not achieved without cost: the small porpoises must obtain three times as much food as a terrestrial mammal of the same weight and must rise to the surface and breathe three times more often than they would if their metabolic rate were terrestrial.

On the basis of the porpoise data we went on to make similar thermal calculations for larger cetaceans to see how much more easily a large mammal, with its more favorable geometry, could keep warm in cold water. We first considered a theoretical whale weighing 1,000 times more than a harbor porpoise; since the two sea mammals are similar in shape, the whale's linear dimensions must increase about tenfold. The theoretical whale therefore has 1,000 times more tissue to generate metabolic heat but the heat passes out to the surrounding water through a surface area only 100 times larger than the porpoise's.

Consider this diminished heat loss in terms of blubber insulation only. Whereas the small porpoise could balance its heat loss with the aid of a blubber layer two centimeters thick, the 1,000-times-heavier whale, given the same metabolic rate as the porpoise, could make do with a blubber layer no more than two millimeters thick. Actually, however, such a whale would have a blubber layer some 20 centimeters thick and even in polar seas would face a cooling problem. Such a large cetacean appears to be 100 times overinsulated.

If the large cetacean's blubber layer is far thicker than considerations of insulation require, what other functions does it serve? First, the fatty tissue, which is low in specific gravity, provides enough buoyancy to offset the negative buoyancy of the whale's muscle and skeleton. Second, the blubber constitutes a food store large enough to take care of the whale's metabolic needs on the animal's seasonal migrations, often across thousands of miles of food-poor waters.

Because of the general rule that relates the metabolic rate of mammals to their size, we would expect a whale to have a rate of heat production per kilogram of tissue considerably lower than the rate we had established experimentally with harbor porpoises. We could see no way to check this expectation by direct measurement of the metabolic rate of a 40-ton animal, and so we were left with estimating the whale's heat production on the basis of anatomical considerations only.

At Steinshamn, a Norwegian coastal whaling station, we dissected out the lungs of a 48-foot fin whale (*Balaenoptera physalus*). We determined the whale's lung volume by first inflating the organs with compressed air and then emptying them through a gas meter. We assumed that the volume of expelled air was equal to the volume of air the whale inhaled with each breath. From observations at sea we were able to determine how often a fin whale breathes. The normal diving pattern for the species is to surface for five quick breaths and then to stay underwater for five minutes; thus in effect a mature fin whale breathes on the average about once a minute.

From the combination of lung volume and breathing rate we could estimate the amount of air a fin whale breathes: 2,000 liters per minute. Now, mammals in general utilize about 6 percent of the oxygen in the air they breathe (21 percent of the total atmospheric mixture of nitrogen, oxygen and other gases). Because each cubic centimeter of oxygen the animal consumes will generate five calories of heat, we were able to estimate, albeit crudely, the fin whale's rate of metabolism. The figure we got was, like that for the harbor porpoise, significantly above the curve for terrestrial mammals. On a weight-for-weight basis, however, the fin whale's metabolic rate is even lower than that of sloths, which have a rate reflecting their lethargic movements.

When one takes into consideration both the large whales' lower absolute metabolic rate and their ability to accumulate a food reserve in the form of blubber, it becomes evident that all large whales are relatively independent of local food supplies. They can satisfy their energy needs for three to six months simply by living off half of their supply of blubber. This leaves them free to feed in the rich polar seas and to give birth to their young thousands of miles away in warm tropical waters.

In comparison the small harbor porpoise leads a much more restricted life. Its higher metabolic rate makes its need for food much more pressing and limits the time available for migration. Moreover, its higher rate makes it more difficult for it to hold its breath: when the porpoise dives to search for food, it must depend on such oxygen as it can take along. Just how the porpoise makes do with what oxygen it has became the focus of our next investigations.

We already knew that the mammals' high-energy way of life had been made possible by a correspondingly larger consumption of oxygen; this in turn depended on a parallel evolutionary development of respiratory and cardiovascular systems that were more competent than those of amphibians and reptiles. When the terrestrial mammals that were ancestral to the cetaceans abandoned the land (with its unlimited access to oxygen) in favor of an aquatic life, a still further evolutionary development favoring greater oxygen storage and transport was called for. How had this been accomplished?

Consider the sperm whale, the cetacean with one of the most impressive of all underwater capabilities. Whereas the porpoise can stay submerged for only a

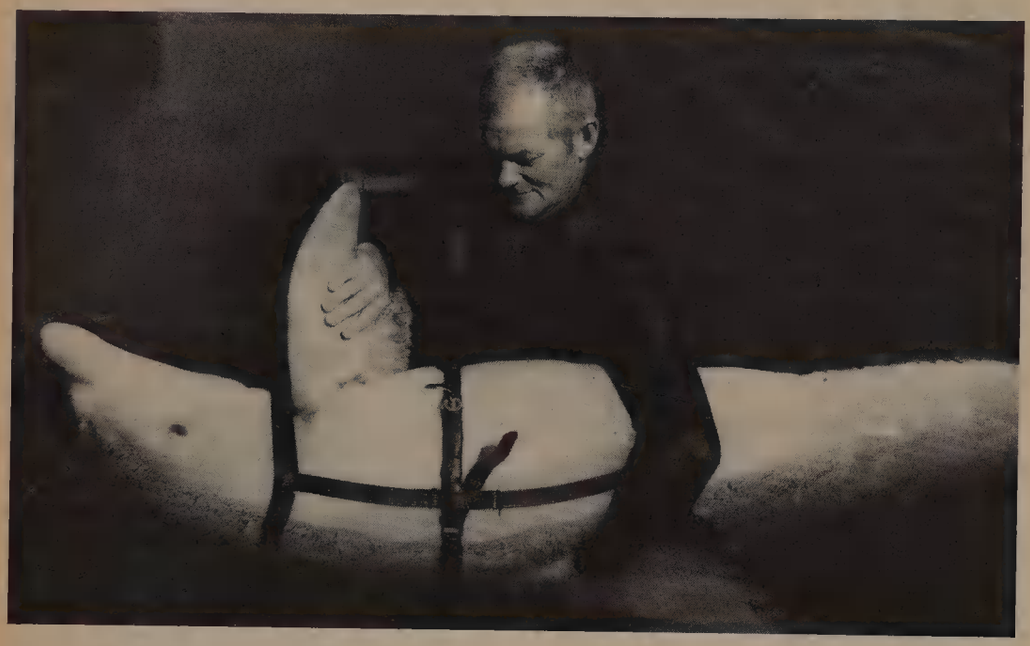
few minutes, sperm-whale dives can last for as long as an hour. Yet even the porpoises' briefer underwater performance puts an unassisted human diver to shame. The answer, for large and small cetaceans alike, is that they can take more oxygen down with them than a human diver can.

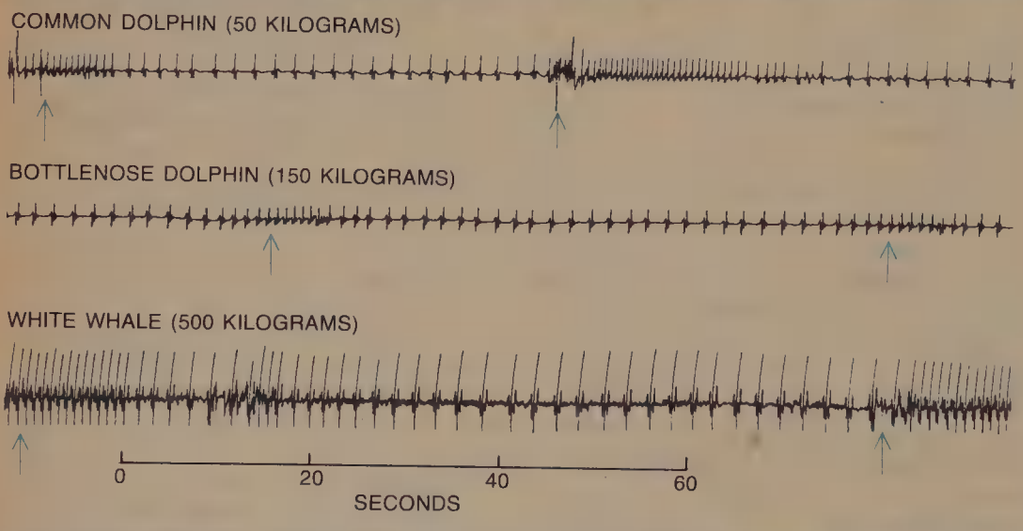
One factor in this performance is the cetaceans' considerably greater volume of blood: up to two or three times more blood per unit of body weight than is the case in human beings. The cetaceans' blood also has a somewhat greater oxygen-carrying capacity than human blood. In addition the whale carries a further store of oxygen in combination with myoglobin in its muscles. The myoglobin (a protein that is actually a subunit of the oxygen-carrying hemoglobin of the red blood cells) is responsible for the muscles' characteristic dark red color. Indeed, in some cetacean species there is enough myoglobin in the muscles to carry more than half again as much oxygen as the animal's red blood cells do.

When a whale comes to the surface after an extended dive has depleted both these oxygen stores, its highly efficient vascular system rapidly passes the blood through the lungs for oxygen replenishment with the aid of a particularly powerful heart and a circulatory system that features large networks of capillaries (*retia mirabilia,* or "wonderful nets") and large sinuses in the venous system. Although the exact function of these unusual structures can only be guessed at, it seems likely that they facilitate the rapid reloading of the red blood cells with oxygen. The fast circulation of the blood is reflected in the fact that while the animal is breathing at the surface it has a higher heart rate. This constitutes a fundamental respiratory difference between cetaceans and terrestrial mammals, whose breathing is usually regular and whose blood flow is relatively smooth.

Meanwhile the recharged red blood cells convey part of their oxygen load to any oxygen-depleted myoglobin in the animal's muscle cells. The transfer is easy because myoglobin has a much lower oxygen-loading tension than hemoglobin. Myoglobin thus binds oxygen more tightly than hemoglobin does, inhibiting it from reentering the circulatory system. It is best to think of the myoglobin-bound oxygen as a purely local oxygen resource, available to the surrounding muscle in the course of a dive. Most of it lies within easy diffusing distance of the muscle cells' mitochondria: the intracellular particles where respiration occurs.

Myoglobin is present in the muscles of terrestrial mammals too, but in much smaller amounts. One of its functions in terrestrial mammals is to serve as a local oxygen source for muscle during contraction, when intramuscular pressure tends to throttle capillary blood flow, cutting off the normal supply of oxygen carried by red blood cells. It is evident that the much more abundant supply of oxygen carried by cetacean myoglobin is another example of a diving adaptation that has arisen through the enhancement of a preexisting mammalian capability.





TOOTHED WHALES' HEARTBEATS WERE RECORDED, by telemetry for bottlenose dolphins such as the one carrying an ultrasonic transmitter in the photograph at the top and directly in the case of the remaining electrocardiograms. These show the heartbeat of a 50-kilogram common dolphin (*Delphinus delphis*) and that of a 500-kilogram white whale (*Delphinapterus leucas*). Arrows in color indicate the start of inhalation in each instance. The heartbeat of all three animals accelerated as they inhaled, an action that would have accelerated oxygen uptake. Thereafter, whether the animals were diving, were resting on the surface or were out of the water altogether, their heartbeat decelerated until a basal rate was reached. Earlier experiments (conducted by the physiologist P. F. Scholander with seals and ducks under restraint) had shown a radical decrease in heart rate following involuntary submergence, but the three cetaceans did not show such a decrease. The authors' findings suggest that the reaction of animals under restraint was one of alarm as they were involuntarily submerged.

When we began our study of cetacean diving capabilities, a widely accepted explanation of the long dive times characteristic of aquatic animals already existed. It derived from experiments done by P. F. Scholander of the Scripps Institution of Oceanography demonstrating that when a restrained animal (a duck or a seal) was forcibly submerged, its heart rate decreased abruptly and the reduced blood flow was shunted in such a way as to largely shut off the circulation of blood to the animal's muscles and viscera. The reduced cardiac output went almost entirely to the brain and the heart, organs that were considered to have a need for a continuous supply of oxygen. (For example, a few seconds' interruption in the flow of blood to the human brain causes unconsciousness.) All that was available to the oxygen-deprived muscles of Scholander's subjects was an anaerobic energy pathway that resulted in an accumulation of lactic acid as a metabolic by-product.

This logical sequence of events has become the standard textbook explanation of the extended diving capacity not

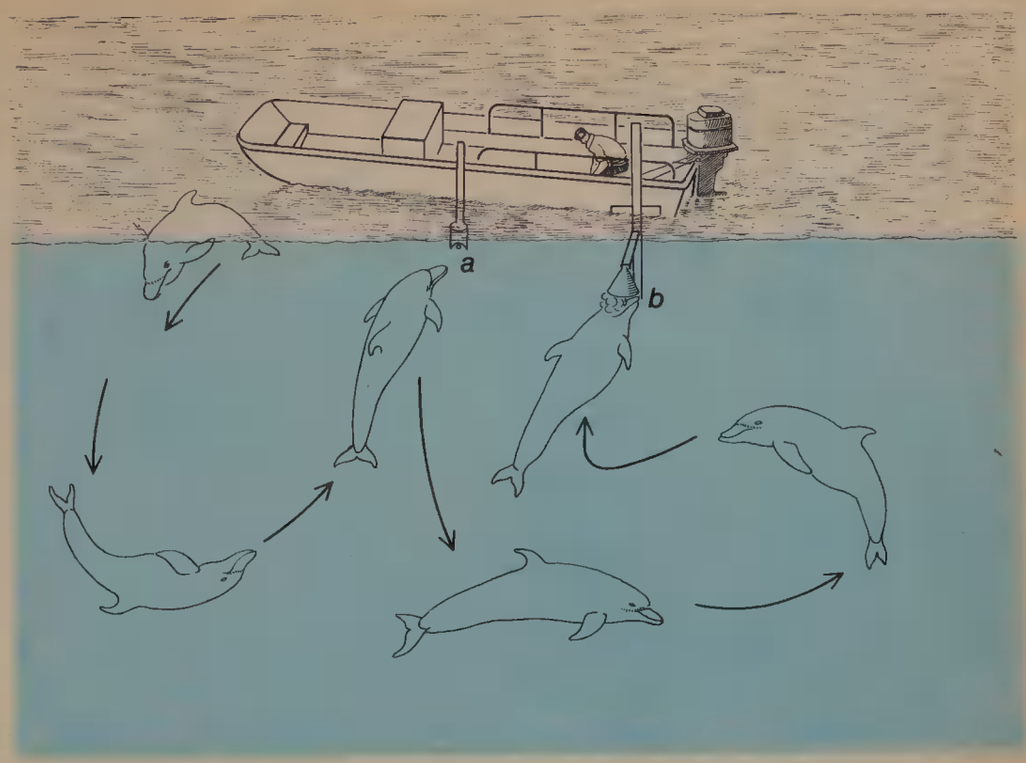
only of ducks and seals but also of other waterfowl and of marine mammals in general. There were, however, virtually no physiological data for cetaceans. Indeed, the one time a restrained porpoise was experimentally deprived of air in this way it went into shock and died. We undertook to collect evidence on this hypothetical diving reflex from free-swimming cetaceans.

As a first step we fitted porpoises with ultrasonic transmitters so that we could monitor the animals' heartbeat as they swam. We were surprised, first of all, to find no evidence of the abruptly diminished heartbeat on diving that was characteristic of forcibly submerged ducks and seals. The porpoises seemed to maintain a relatively high level of general circulation throughout their dive. We therefore set out to find a more direct measure of how quickly the porpoises depleted their store of oxygen while they were submerged.

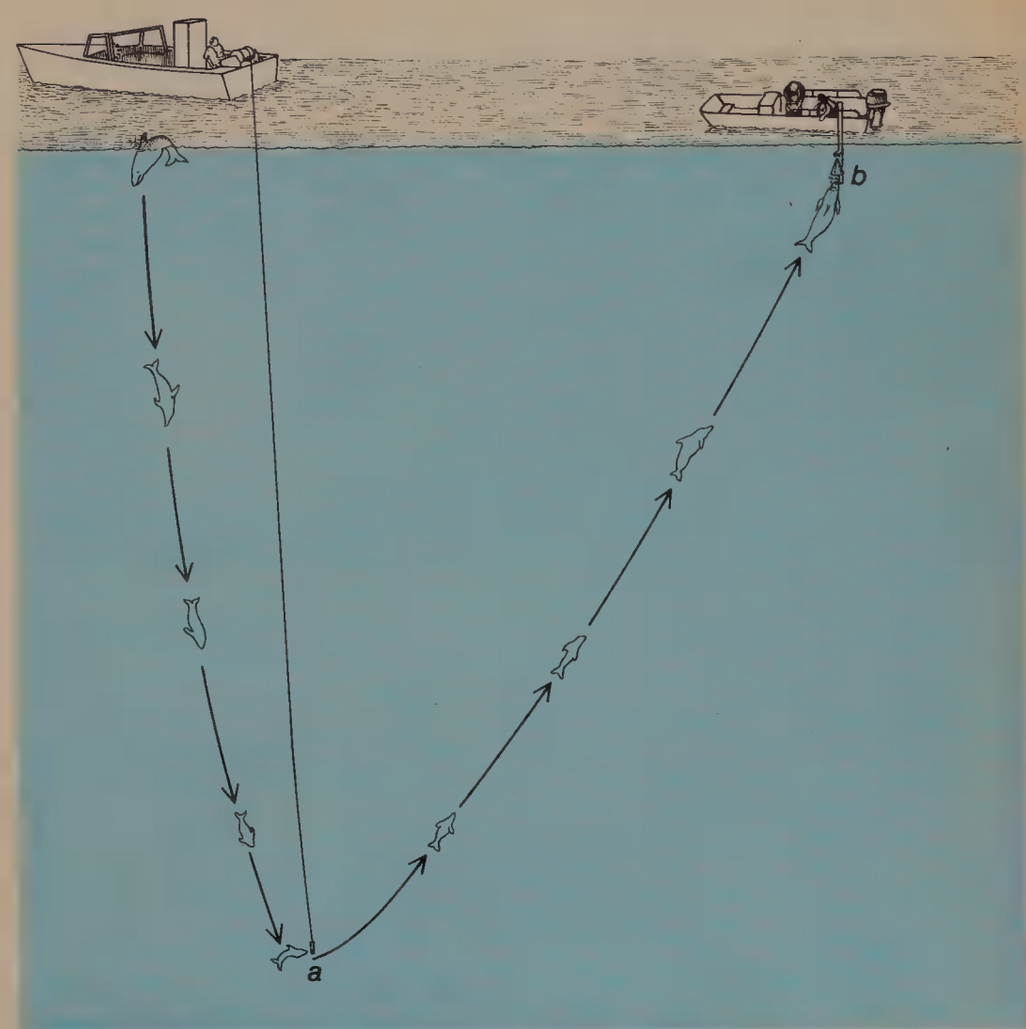
Fortunately for us the U.S. Navy at that time was training bottlenose dolphins (*Tursiops truncatus*) to act as messengers for deep-sea divers. One Navy trainer, William Scronce, undertook to train one of the dolphins, named Tuffy, to exhale on command into an underwater funnel so that we could collect the expired breath for analysis. An underwater buzzer was Tuffy's signal to perform. If he pushed his snout against the buzzer and exhaled into the funnel, he received three fish as a reward. Another buzzer served as a signal for Tuffy to begin his dive. By varying the interval between the sounding of the "dive" buzzer and the "collection" buzzer we could obtain air samples over an entire range of breath-holding times. Tuffy could easily hold his breath for as long as six minutes, but after a much shorter time he grew increasingly impatient as he waited for the "collection" buzzer to sound; looking up through the water at us, he would even gnash his teeth.

Studying the data Tuffy provided us, we were surprised to see how quickly the oxygen concentration in the porpoise's lungs dropped off. For example, after only three minutes the gas in Tuffy's lungs had fallen from its normal atmospheric level of 21 percent oxygen to less than 2 percent. For the remaining three minutes of these "easy" dives, therefore, he was virtually anaerobic. After such a breath-hold he would not perform again until he had spent four or five minutes breathing rapidly on the surface, obviously repaying the oxygen debt he had just incurred. After dives of two minutes or less, however, he would readily dive again following only a brief breathing interval at the surface.

We could now understand why we had failed to observe an abruptly diminished heart rate when our ultrasonically equipped porpoises dived. Like Tuffy, they had made only a partial effort, if



NAVY-TRAINED DOLPHIN TUFFY was taught as is shown here to submerge when one buzzer (*a*) sounded and to stay underwater until a second buzzer (*b*) sounded. Tuffy then swam to the second buzzer, touched it with his snout and exhaled into a funnel. After he exhaled he received a reward of three fish. The study showed that Tuffy could remain underwater comfortably for six minutes but had exhausted almost all his oxygen supply within three minutes.



DEEP-DIVING ABILITY OF THE DOLPHIN was tested in a second experiment. First a buzzer and a switch (*a*) were lowered to a depth of 150 meters. Turning the buzzer on was Tuffy's signal to dive and switch it off. The dolphin then returned to the dinghy, to touch a second buzzer (*b*) and claim his reward. Tuffy had no difficulty with 150-meter dives but needed time at the surface to recover after turning off the "dive" buzzer at a depth of 300 meters.

The Search for Alexander

The son of a king, a pupil of Aristotle, an awesome warrior and brilliant military strategist, he was blessed with a vision of a world united and civilized.

His achievements inspired artists and thinkers for over four hundred years. This collection forms a part of his legacy. Gold statuary and jewels. Silver vessels and decorative ornaments. Bronze objects. Terra cotta and marble statues. Important discoveries. Yet the most important discovery eludes us: his final resting place. Alexander the Great. Come join the search at the newly renovated Royal Ontario Museum in Toronto, March 4 – July 10. Admission: $5.00 for adults; seniors and students, $4.00; children, 12 and under, $3.00.

For more information, call Ontario Travel TOLL FREE: **1-800-828-8585;** in New York State **1-800-462-8404.**

ONTARIO yours to discover!®

any at all, to conserve their initial store of oxygen. That store was quickly consumed; thereafter an oxygen debt developed in all their tissues, including the brain. We found that all the aquatic air breathers we observed (diving birds, seals and manatees in addition to our cetaceans) tried to avoid such an oxygen debt by limiting their dives to short periods in order not to exhaust their supply of oxygen. For example, when Tuffy and his dolphin companions swam without instructions to the contrary, dives lasting for more than 2.5 minutes were the exception rather than the rule. Submerging for shorter intervals, however, the dolphins might continue diving for hours.

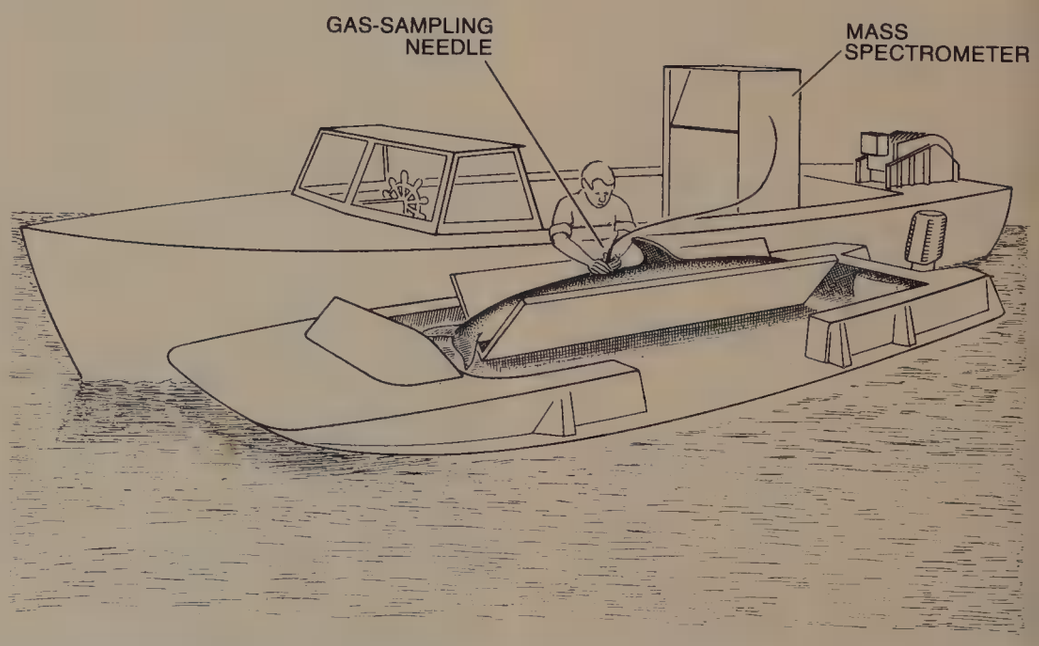
When a human diver has spent enough time underwater for a substantial amount of nitrogen to go into solution in his blood and then he comes up too rapidly, the nitrogen forms tiny bubbles that can cause painful symptoms ("the bends") and even death. How did our porpoises, let alone the large whales that may dive deeper than a kilometer for more than an hour, avoid the bends? With this question and related ones in mind we arranged some deep-diving experiments for our trained bottlenose dolphin.

One buzzer, controlled from the surface, we lowered on a line to one or another of a variety of chosen depths. Tuffy was trained to swim down to the buzzer when we sounded it and turn it off by pushing a switch with his snout. When the animal turned off this "dive" buzzer, Scronce immediately sounded the "collection" buzzer at the surface. Tuffy then quickly swam up from his dive to the "collection" buzzer and exhaled into the collection funnel.

The scientific literature on porpoises suggests they do not dive very deep. When we began our experiments, we lowered the "dive" buzzer to a depth of 150 meters. Tuffy easily managed trips to this depth, came up to give us his gas sample and sometimes demanded all three of his reward fish before even bothering to inhale; he was clearly far from his depth limit. When we lowered the "dive" buzzer to 300 meters, we found that Tuffy needed a much longer recovery time. He was obviously approaching his limit. (Navy trainers in Hawaii have since trained a related species to descend to nearly 600 meters.)

Tuffy seemed to be untroubled by the increasing water pressure at greater depths (as much as 30 atmospheres, or 440 pounds per square inch, at a depth of 300 meters). We expected that at such a high external pressure the animal's lungs would be almost completely collapsed. By rigging the "dive" buzzer with a flash camera we were able to record Tuffy's appearance as he turned off the buzzer 300 meters below the surface. We found that his rib cage was indeed much deformed by the hydrostatic pressure pushing on his lungs.

We next tried to give another trained dolphin the bends by directing it to dive repeatedly. As soon as its surface breathing indicated its recovery from the previous dive we would send it down



THIRD EXPERIMENT was aimed at discovering whether a bottlenose dolphin is susceptible to "the bends" (the painful symptoms that can result from the absorption of atmospheric nitrogen by the blood in the course of a long dive). On signal the dolphin dived repeatedly to a depth of 100 meters, spending more than 20 minutes out of an hour at an average depth of 50 meters. The animal then swam onto a floating couch next to a boat containing a mass spectrometer; the experimenters pushed a gas-sampling needle through its blubber layer and into its muscle. The dolphin's muscle proved to be saturated with nitrogen to an extent dangerous to a human diver. The authors attribute the dolphin's freedom from the bends to a diving adaptation.

again, until in the course of an hour it had remained below 50 meters for more than 20 minutes. The dolphin still showed no apparent sign of distress. Then, with the help of Robert S. ("Red") Howard of the Scripps Institution, we measured the actual amount of nitrogen accumulated in the animal's muscles as the result of these many consecutive submergences. It was no easy task. First the dolphin was trained after such a sequence of dives to swim onto a floating couch alongside our boat. That enabled us to insert a small gas-sampling needle through its blubber layer into the underlying muscle. A tube led from the needle to a mass spectrometer (aboard a second boat) that could directly measure the concentration of nitrogen dissolved in the muscle. Doing all this while at sea in small boats makes a substantial demand on the experimenter's patience.

Again we got a surprise. The concentration of dissolved nitrogen in the dolphin's muscle tissue was indicative of a degree of supersaturation that in a human diver would have been dangerous. We could only conclude cetaceans can tolerate nitrogen concentrations that would give a human diver the bends. How they do so we cannot yet say, but the capability is, of course, what one should probably expect of a well-adapted order of deep-diving mammals.

What has the evolution of this superior diving capacity meant to the ecology of the world's oceans? Some aquatic mammals, for example the sea otter and the manatee, have evolved to more or less nibble at the edges of the oceanic food chain. Others, such as the carnivorous suborder of pinnipeds (seals, sea lions and walruses), are more at home in the open ocean but still need to go ashore for mating and reproduction. Only cetaceans have evolved to deal with a fully oceanic way of life. What they gained thereby is a vast food supply in the form of other marine animal life. The whales and porpoises have established themselves as harvesters at the very top of the marine food chain.

The success of these physiologically adept marine predators may be judged by their numbers. Many species of cetaceans are commonly thought to be in imminent danger of extinction; this conclusion is based largely on a misperception of their current population levels. For example, the total number of blue whales is greater by at least one order of magnitude than the total number of such endangered terrestrial species as the orangutan or the giant panda. In fact, most cetacean species exist in substantial numbers, and some populations are even reaching the level where they serve as entertainment. Whale watching is a growing winter sport on the California coast, and in the Atlantic excursion boats putting out from the Boston area

Better Than Jogging, Swimming or Cycling

NordicTrack

Jarless Total Body Cardiovascular Exerciser
Duplicates X-C Skiing For The Best Way To Fitness

NordicTrack duplicates the smooth rhythmic total body motion of XC Skiing. Recognized by health authorities as the most effective fitness building exercise available. Uniformly exercises more muscles than jogging, swimming, cycling and rowing.

Does Not cause joint or back problems as in jogging. Highly effective for weight control and muscle toning.

Easily Adjustable for arm resistance, leg resistance and body height. Smooth, quiet action. Folds compactly to require only 15 by 17 inches of storage area. Lifetime quality.

Used In thousands of homes and many major health clubs, universities, and corporate fitness centers.

Call or Write for FREE BROCHURE
Toll Free 1-800-328-5888 MN 612-448-6987
PSI 124 F Columbia Crt., Chaska, Minn. 55318

GEIGER - COUNTER for EVERYONE

The MINI-Geiger-Counter "RAYtronic" fits in every breastpocket and can be employed as a warning or measuring set.

Special tubes for different ionizing radiations can be quickly exchanged.

The optical and acoustical signals can be complemented by a Digital-Reading by attaching a modified scientific calculator which can count and accumulate the impulses.

Exact measuring series are thus possible outside the laboratory.

Complete set: RAYtronic-Body with special measuring tube for ALPHA-, BETA- and GAMMA-rays (with mica window) and attachable calculator = US-$ 343.-

U.S.A.:
GermanTronic, 195, Kings Highway, Woodmont, Conn. 06460

Europe:
INNOVA GmbH, Talstr. 39, D-6057 Dietzenbach

—1 year of guarantee!—

A BERLITZ® LANGUAGE COURSE in a handsome briefcase for only $125 if you pay for it BUT ...wouldn't your company be interested in having you speak a second language?

These days a second language can help you help your company and so help you earn more. The many hispanic nationals now in our work force respond more readily to someone who speaks their language. Also, selling today is international. Whatever language it would pay you to speak, Berlitz has a do-it-yourself course in a handsome briefcase you can take anywhere. Contents—90 minute introductory tape, 40 lessons on five 60-minute cassettes, 6 illustrated word and phrase books, rotary verb finder to help you find verb tenses instantly. If not satisfied, return within 10 days and get your money back. Learn from the leader!

For your convenience on credit card orders dial toll-free
1-800-228-2028 ext. 35
and **refer to Dept.3111**
24 hours a day, 7 days a week.

Why not give us a ring—right now! (our expense)

(Nebraska residents dial 402-571-4900)

Berlitz Publications, Inc., Dept.3111
3490 Lawson Blvd., Oceanside, NY 11572

Send **Berlitz Comprehensive Cassette Course(s)** checked. $125 each in briefcase plus $4 for shipping and insured delivery.

☐ **French** 86100 ☐ **German** 86101 ☐ **Italian** 86102 ☐ **Spanish** 86103

Enclosed check _____ money order _____ payable to Berlitz.
Or charge my AMEX _____ Diners Club _____ VISA _____
MasterCard Interbank # _____ Exp. Date _____
Card # _____ Exp. Date _____
Name _____
Address _____
City _____ State _____ Zip _____

N.Y. residents add sales tax. Allow 4 weeks for delivery.

WHERE DO BLACK DOCTORS COME FROM?

Nearly half of all black doctors practicing in the U.S. today are graduates of Meharry.

Sure, there are black medical students at Harvard and Stanford and Johns Hopkins. But no single private medical school in the U.S. produces as many black physicians as Meharry Medical College.

Ninety percent of our students require financial aid to some degree. To make more funds available for scholarships and student loans, and to maintain the quality of our training, our goal is to raise $27 million.

Meharry must depend largely on private and corporate contributions. We are eligible for only limited funding under Federal and State programs—and, as a professional school, receive no resources from the United Negro College Fund.

Won't you help? Because without Meharry, a lot of qualified youngsters couldn't even consider a medical career.

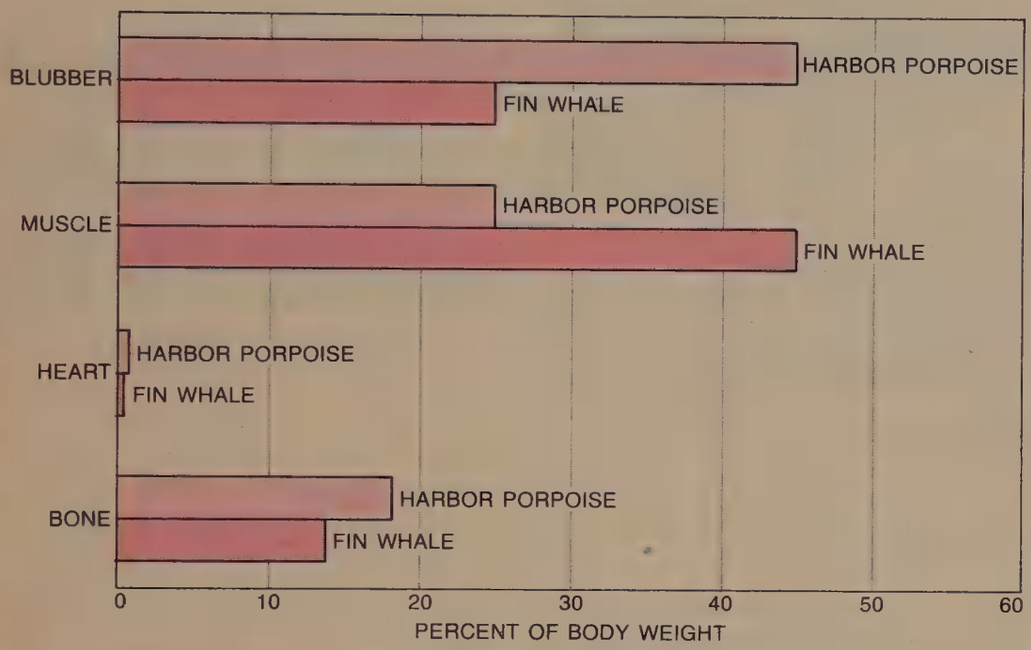
OPEN WIDE FOR MEHARRY
MEDICAL COLLEGE

Send your tax-deductible contribution, in any amount, to **Meharry Medical College, Second Century Fund, Nashville, TN 37208.**
The need is urgent, and your support is appreciated.

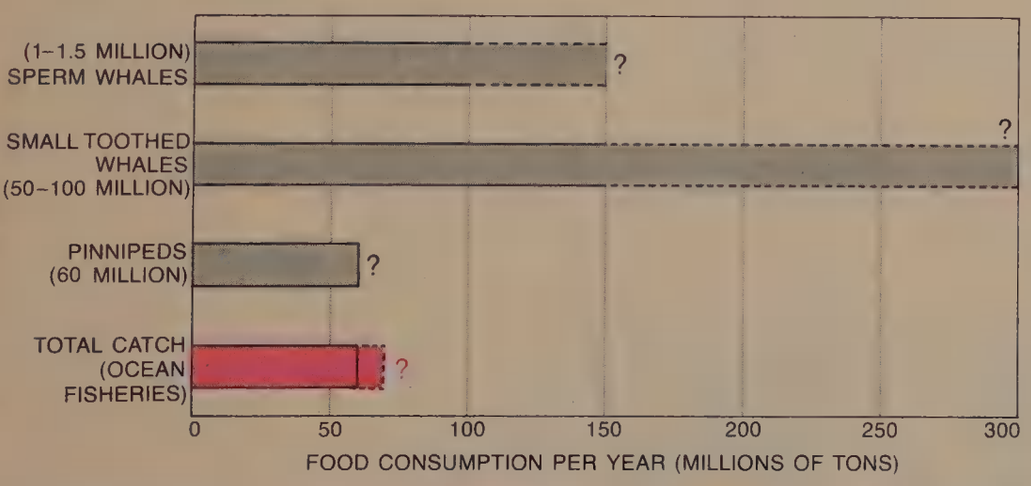
assure tourists of frequent whale sightings after a trip of only an hour or two.

Sperm whales were probably twice as numerous in the generations before modern whaling began than they are today, but even so their modern number is far from low. The British cetologist Malcolm R. Clarke reckons the present population of sperm whales at about a million (others say 1.5 million) and estimates that they annually consume some 100 million tons of deep-sea squid, an oceanic resource that is otherwise scarcely touched. It is instructive to compare the predation of this single cetacean species with the total of the world's annual fishery yields: 60 to 70 million tons.

An annual-consumption estimate for the next-ranking aquatic mammals, the pinnipeds, has been made by Michael E. Q. Pilson of the University of Rhode Island. Pilson's estimate has the advantage of better-known population numbers, since pinnipeds must come ashore for breeding and can be more accurately counted than the 70 cetacean species. Censuses have been taken at most pinniped breeding areas, and their annual food needs are known from the feeding of captive species. Pilson estimates that the seals, sea lions and walruses consume 60 million tons of fish per year, or approximately the amount harvested by man. Extrapolating from the number of porpoises killed annually by the opera-

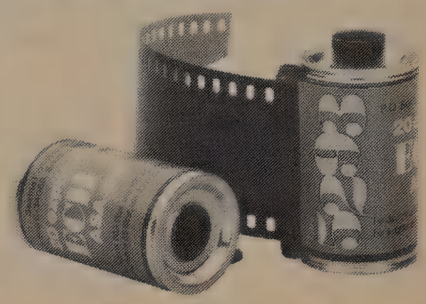


DIFFERING PROPORTIONS OF TOTAL BODY WEIGHT devoted to various components are shown in this graph. A small odontocete, the harbor porpoise, is compared with a large mysticete, the fin whale of cold polar waters. Only about a fourth of the porpoise's weight consists of muscle, whereas nearly half of its weight consists of blubber. The geometry of surface-to-volume ratios provides the fin whale with a thick blubber sheath accounting for only a fourth of its total weight, whereas nearly half of the total consists of its muscle tissues.



ECOLOGICAL ROLE OF MARINE MAMMALS appears to be at least as significant as that of the land mammals. The graph compares the marine mammals' estimated "harvest" of other marine organisms with the total recorded international catch of fish. A million sperm whales consume some 100 million tons of squid each year (more if sperm whales are actually more numerous). The 50 to 100 million porpoises and other small odontocetes consume between 100 and 300 million tons of fish. Estimates of consumption by whales other than the sperm are less reliable, but seals, sea lions and walruses consume 60 million tons of fish annually. The total catch by marine fishermen worldwide is somewhere between 60 and 70 million tons (*color*).

35mm Color



Prints and Slides from the same roll

ECH II 5293® . . . Eastman Kodak's professional color motion picture film . . . adapted for still use in 35mm cameras by Seattle FilmWorks. With micro-fine grain and rich color saturation, its quality meets the exacting standards of Hollywood filmmakers. You get remarkable versatility: shoot in **low light or bright light at up to 800 ASA;** use it indoors or out without a filter; and you have the option of **prints or slides, or both, from the same roll.** Enjoy the very latest in photographic technology with substantial savings.

©1983 Seattle FilmWorks

INTRODUCTORY OFFER

☐ Rush me 2 rolls of 20-exposure Kodak 5293®. Enclosed is $2.00. I'd like to be able to get color prints or slides (or both) from the same roll and experience the remarkable versatility of this professional quality film. Offer limited to 2 rolls per customer.

NAME_____

ADDRESS_____

CITY_____

STATE_____ ZIP_____

909

Mail to:
Seattle FilmWorks, 500 Third Ave. W.
P.O. Box C-34056, Seattle, WA 98124

Another QUESTAR® . . .totally new. . . the long-range microscope

This is the QM1, the new instrument by Questar Corporation. It focuses at distances from 22 to 77 inches from the target. With it you will inspect hot materials at safe distances, observe live specimens in natural habitat, inspect micro-circuits, use it for crystal growth studies and a thousand other applications. It will let you scrutinize and photograph living organisms or radio-active samples with no damage to either one or yourself. And the magnificent design of these optics permits resolution below 2.5 microns. With a selection of eyepieces and extensions, magnifications can range as high as 150X.

Shown below on the right is a chip from a micro-circuit photographed with a normal 50mm lens to show its tiny size. On the left is a QM1 photograph of the lower right corner, reproduced from an 8 × 10 enlargement of 35mm Kodak Technical Pan. Distance 22 inches.

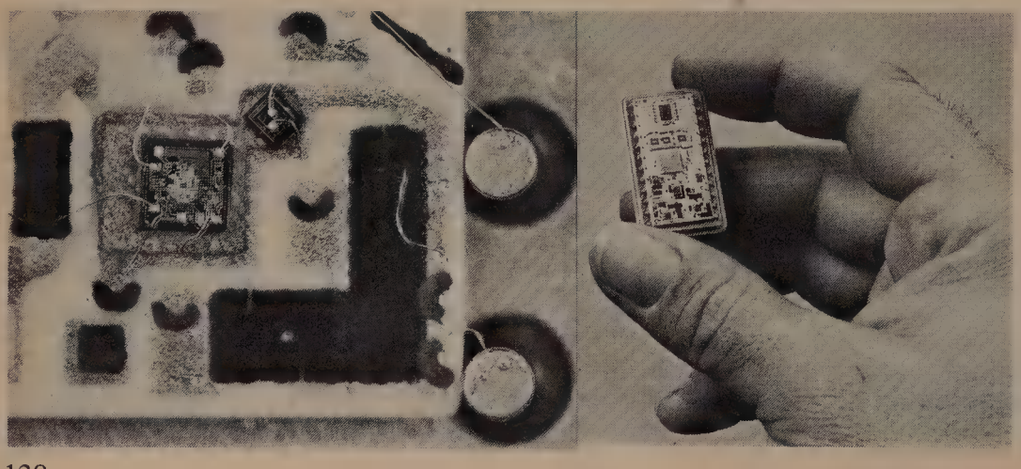
With the Questar QM1 we can provide full support systems of equipment and accessories, including video and high-speed motion cameras, lighting equipment, micro measurement and image analysis, and computer storage and display units. These systems are individually assembled to meet the customer's requirements.

NEW LITERATURE WITH SPECIFICATION AND PHOTOGRAPHS OF QM1 PERFORMANCE ON REQUEST.

The QM1: *non-destructive testing, non-disruptive observation, non-contact gauging.*

QUESTAR

P.O. Box C, Dept. 205, New Hope, PA 18938
(215) 862-5277



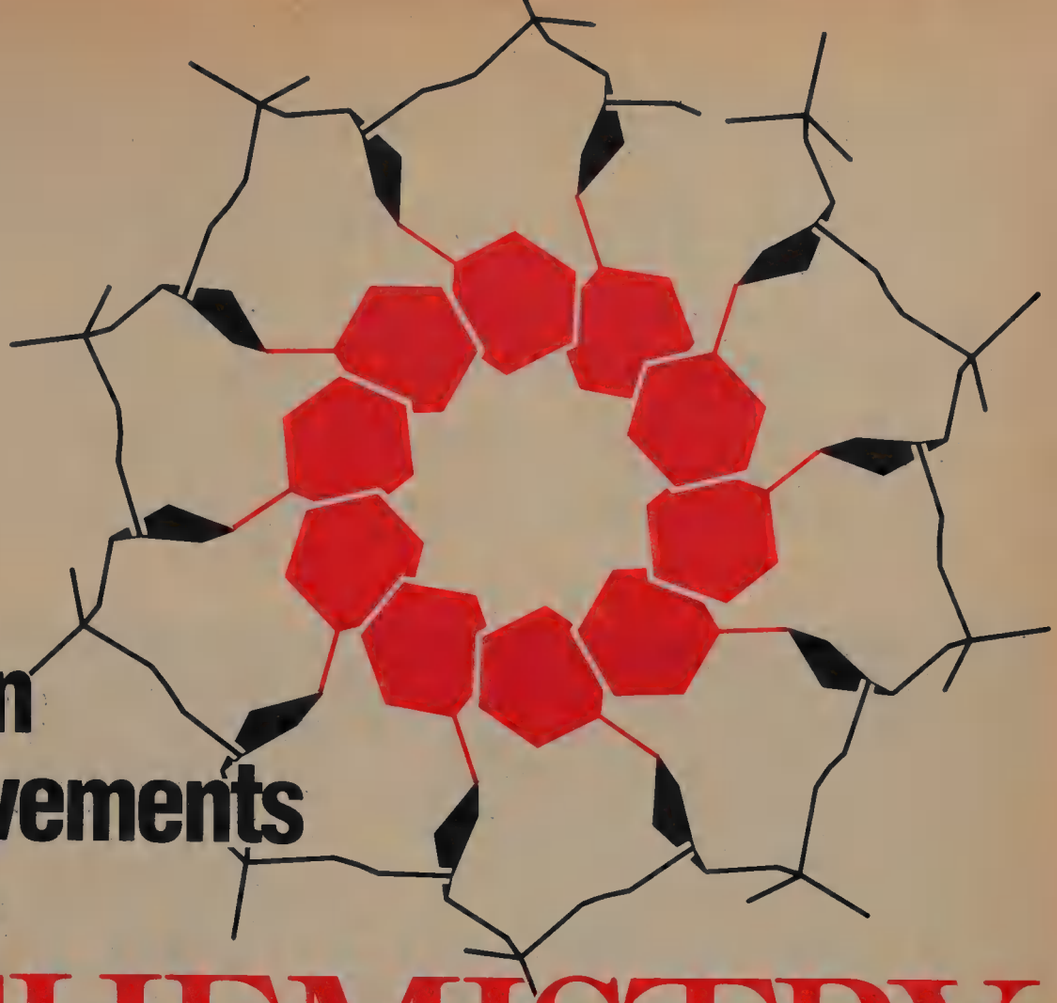
tions of the tuna fishery in the eastern Pacific, the worldwide stock of porpoises and dolphins may approach 100 million. Even if the population is only half this size, its annual food consumption must total some 150 million tons.

Such large numbers may come as a surprise, and we hasten to add that they must be considered tentative. Indeed, unless oceanic animals are harvested by man either accidentally or deliberately, the size of their populations is not likely to be well known. Two examples will suffice. Until recently Fraser's dolphin was known only from a single skeleton collected in 1954. Now that people are aware of the species, Fraser's dolphins have been sighted in schools numbering in the hundreds in Philippine waters, and the animals appear to be widely distributed in the tropical zones of the Pacific and Indian oceans. The North Pacific population of Dall's porpoise was estimated at only a few tens of thousands. Since it became known that Japanese tuna fishermen kill as many as 10,000 Dall's porpoises every year, new surveys have raised the population estimate to a million or more.

Thus it is evident that the cetaceans' role in marine food chains is far greater than man's. This does not mean that man's overfishing has not heavily depleted local fish stocks, as it has the stocks of some whale species, particularly in the Antarctic. It does indicate, however, that human harvesting, being largely confined to a few highly productive areas, has probably not seriously affected the gross nutrient dynamics of most ocean areas.

One feature of gross nutrient dynamics is upwelling: the lifting of nutrients from deep waters such as that accomplished by the rising currents off southern Africa in the Atlantic and off South America in the Pacific. Cetaceans play an analogous role on a surprisingly large scale. All other life in the sea tends to move nutrients downward. Tied to the surface by their air-breathing, marine mammals tend to move nutrients upward. Even the whales' fecal output does not move downward: because it is liquid, it tends to disperse rather than sink when it is released.

To those of us who are interested in the whales' well-being for emotional as well as biological reasons there is new cause for hope. Hunting these animals in distant seas is an energy-intensive activity. As the commercially desirable baleen whales have become harder to find, the harvesters' costs, largely in fuel consumed, have risen to a point where economics will increasingly discourage the activity. Ironically the Organization of Petroleum Exporting Countries (OPEC) may in the end do more than conservationists can to relieve hunting pressure on these remarkable marine mammals.



A celebration of the achievements of a science

BIOCHEMISTRY

SECOND EDITION

LUBERT STRYER

Stanford University

949 pages
983 illustrations
$32.95

Lubert Stryer's *Biochemistry* has become the standard against which all other biochemistry books should be measured. It has shaped and continues to shape the subject it teaches.

First published eight years ago, *Biochemistry* immediately set new standards in the teaching and presentation of biochemistry. Emphasizing the process of discovery, Stryer aimed "to provide an introduction to the principles of biochemistry that gives the reader a command of its concepts and language." Reviewers commented that it was "doubtful whether there will be another text which will appeal to the student as strongly as this one does" (*Nature*).

The Second Edition, "bigger and even more splendid," is "a most handsome and thought-provoking new edition of a fine textbook" (*TIBS*). "It continues the tradition of excellence established in his first edition" (*American Scientist*), brilliantly achieving the author's aim as stated in the Preface: "to weave new knowledge into the fabric of the text...to enhance the book's teaching effectiveness by centering the exposition of new material on common themes wherever feasible and by citing recurrent motifs....and to convey a sense of the intellectual power and beauty of the discipline of biochemistry."

With a clarity and conciseness achieved through the unified narrative of a single author, "....the brevity of the text often belies its depth: the author has a unique ability to expound difficult concepts in a few words." (*Nature*)

Stryer's *Biochemistry,* Second Edition—giving up-to-date coverage of new developments while communicating the excitement and progress of biochemical frontiers.

If you are not using Stryer, compare *Biochemistry* with your present text. After the comparison, even *more* the choice.

W. H. Freeman and Company
41 Madison Avenue
New York, N.Y. 10010

Georg Cantor and the Origins of Transfinite Set Theory

How large is an infinite set? Cantor demonstrated that there is a hierarchy of infinities, each one "larger" than the preceding one. His set theory is one of the cornerstones of mathematics

by Joseph W. Dauben

The nature of the infinite has always been a controversial topic. In antiquity it appeared in the famous paradoxes of Zeno of Elea, who argued with discomfiting clarity that motion is impossible because it requires an object to pass through an infinite number of points in a finite time. The success of Newtonian physics is in large part a result of Newton's introduction of the calculus of infinitesimal rates of change, yet no mathematically rigorous formulation of this awkward but effective idea was offered for more than 200 years. In modern times problems associated with the infinite have appeared in the abstract theory of sets, a theory that provides a foundation for virtually all contemporary mathematics. Moreover, throughout history the idea of the infinite has had theological overtones that have played a role in the acceptance or rejection of the concept and the mathematical and philosophical doctrines associated with it. All these currents of thought converge in the life and work of the German mathematician Georg Cantor.

The substance of Cantor's lifework is well known: in developing what he called the arithmetic of transfinite numbers he gave mathematical content to the idea of the actual infinite. In so doing he laid the groundwork for abstract set theory and made significant contributions to the foundations of the calculus and to the analysis of the continuum of real numbers. Cantor's most remarkable achievement was to show, in a mathematically rigorous way, that the concept of infinity is not an undifferentiated concept. Not all infinite sets are the same size; consequently infinite sets can be compared with one another. For example, the set of points on a line and the set of all fractions are both infinite sets. Cantor was able to prove, in a well-defined sense, that the first set is greater in size than the second. So shocking and counterintuitive were Cantor's ideas to his contemporaries that the eminent French mathematician Henri Poincaré condemned the theory of transfinite numbers as "a disease" from which mathematics would someday be cured. Leopold Kronecker, one of Cantor's teachers and a formidable member of the German mathematics establishment, even attacked Cantor personally, calling him a "scientific charlatan," a "renegade" and a "corrupter of youth."

It is also well known that Cantor suffered throughout his life from a series of "nervous breakdowns," which became increasingly frequent and debilitating as he got older. The breakdowns were probably the symptoms of organic mental disease; a recent study by the British historian of mathematics Ivor Grattan-Guinness, who relies on the evaluation of Cantor's medical records by psychologists at the Halle Nervenklinik (the mental hospital in the East German city of Halle), suggests that Cantor was a victim of manic-depressive psychosis. Nevertheless, it was all too easy for Cantor's early biographers to present Cantor, who was trying to defend his complex theory but was suffering increasingly long periods of mental breakdown, as the hapless victim of the persecutions of his contemporaries.

Such accounts distort the truth by trivializing the genuine intellectual concerns that motivated some of the most thoughtful contemporary opposition to Cantor's theory. They also fail to credit the power and scope of the defense Cantor offered for his ideas. At first he resisted the transfinite numbers because he believed the idea of the actual infinite could not be consistently formulated and so had no place in rigorous mathematics. Nevertheless, by his own account he soon overcame his "prejudice" regarding the transfinite numbers because he found they were indispensable for the further development of his mathematics. Because of his own early doubts he was able to anticipate opposition from diverse quarters, which he attempted to meet with philosophical and theological arguments as well as mathematical ones. Moreover, when he was called on to respond to his critics, he was able to muster his ideas with considerable force. His mental illness, far from playing an entirely negative role, may well have contributed in its manic phases to the energy and single-mindedness with which he promoted his theory.

Georg Ferdinand Ludwig Philipp Cantor was born on March 3, 1845, in St. Petersburg (now Leningrad). His mother, Maria Anna Böhm, came from a family that included several gifted musicians; most notable among them was her uncle Joseph Böhm, the director of a conservatory in Vienna and the founder of a school for violinists that trained many virtuosos of the time. Georg's father, Georg Woldemar Cantor, was a successful tradesman and a devout Lutheran who communicated deep religious convictions to his son. According to Eric Temple Bell's widely read book *Men of Mathematics,* first published in 1937, the younger Georg's insecurities in later life stemmed from a ruinous Freudian conflict with his father, but surviving letters and other evidence concerning their relationship indicate quite the contrary. Georg's father appears to have been a sensitive man who was attentive to his children and took a special but not coercive interest in the welfare and education of his eldest son.

When the young Cantor was still a child the family moved from Russia to Germany, and it was there he began to study mathematics. In 1868 he received his doctorate from the University of Berlin for a dissertation on the theory of numbers; two years later he accepted a position as privatdocent, or instructor, at the University of Halle, a respected institution but not as prestigious for mathematics as the universities at Göttingen or Berlin. One of his colleagues at Halle, Heinrich Eduard Heine, was then

working on the theory of trigonometric series, and Heine encouraged Cantor to take up the difficult problem of the uniqueness of such series. In 1872, when Cantor was 27, he published a paper that included a very general solution to the problem, together with the seeds of what was later to become the theory of transfinite sets.

The problem suggested by Heine arose from the work of the French mathematician Jean Baptiste Joseph Fourier. In 1822 Fourier had shown that the graph of any "reasonably smooth" curve (that is, a curve that exhibits at most a finite number of points of discontinuity) can be represented throughout an interval as the sum of an infinite trigonometric series. In other words, by superposing an infinite number of sine and cosine waves on one another every point on the "reasonably smooth" curve except the points of discontinuity can be approximated to any desired degree of accuracy [*see illustration on next page*]. Such a series is said to converge to the curve or function except at a finite number of points, or "almost everywhere." Fourier's result was of signal mathematical importance because it suggested that certain complicated functions could be represented, or replaced, by a sum of sines or cosines, which is much easier to manipulate mathematically. In order to justify such replacement, however, some guarantee was required that only one such trigonometric series converges to the function. Cantor began to investigate the conditions under which a trigonometric series that converges to a function is unique.

In 1870 Cantor achieved his first result: If a function is continuous throughout an interval, its representation by a trigonometric series is unique. His next step was to relax the requirement that the function be continuous throughout the interval. Suppose, for example, the function to be approximated by the series is defined as follows: The graph of

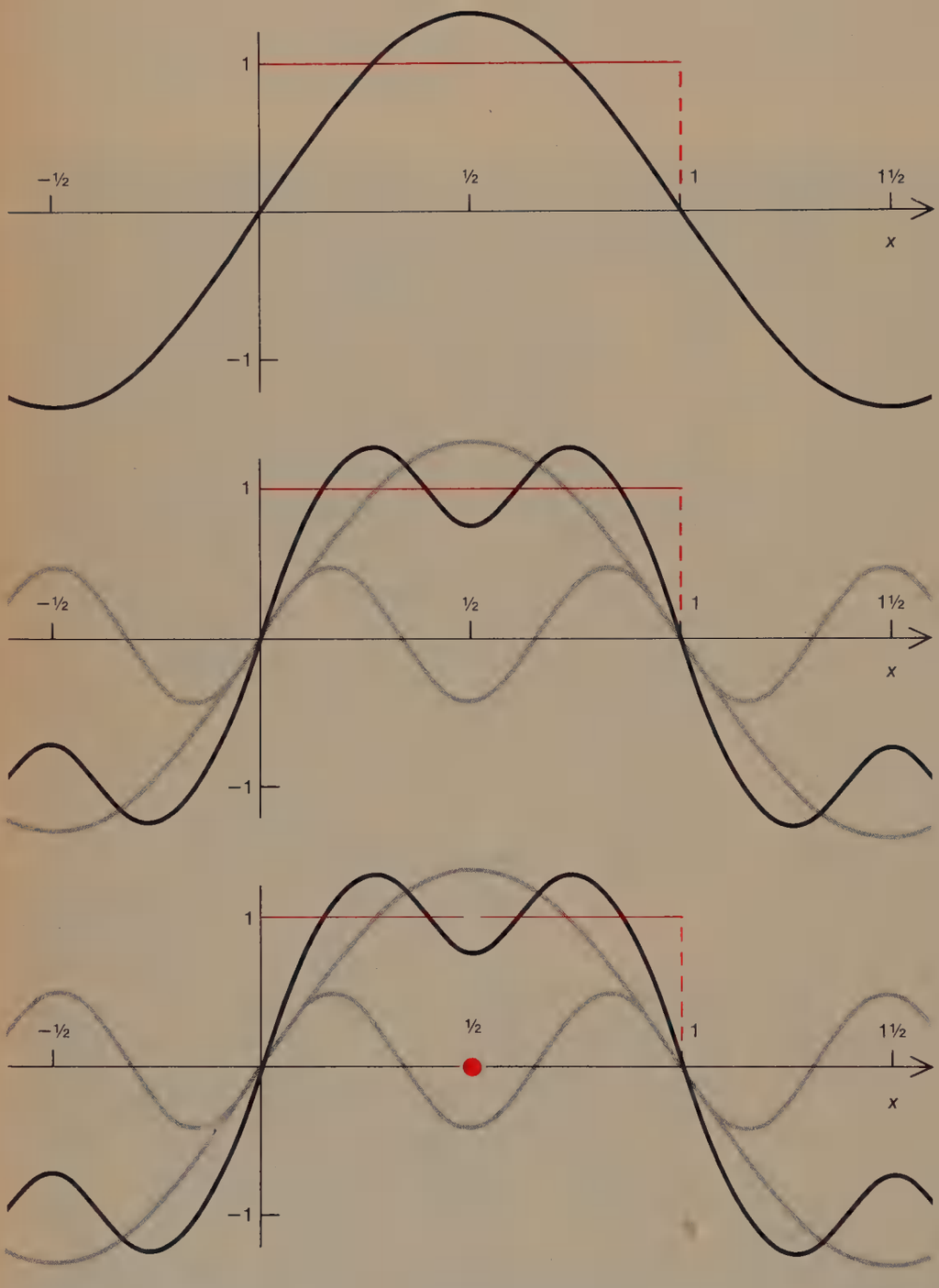
PORTRAIT OF CANTOR and his wife was made in about 1880, when Cantor was at the height of his career. He had already begun to upset the German mathematics establishment by showing that the infinite set of real numbers, represented by the continuum of points on a line, is larger than the infinite set of all fractions. He also showed that infinite quantities called transfinite numbers could be defined that describe such differences. A few years after this photograph was made Cantor suffered a major attack of manic-depressive psychosis, which ultimately put an end to his creative work in mathematics. The original photograph is in the private collection of Egbert Schneider.

the function is a straight line parallel to the x, or horizontal, axis of the graph except at the point on the x axis corresponding to 1/2. For the point 1/2 on the x axis the function assumes the value 0 instead of 1. Cantor was able to show that if convergence is given up at the point where x is equal to 1/2, there is still a unique trigonometric series that converges to the function everywhere else. No other trigonometric series of a similar form exists that also approximates the function. It was then a simple matter for Cantor to extend his result to include all functions having any finite number of points of discontinuity, which Cantor called exceptional points.

In 1872 Cantor's search for a more general statement of his uniqueness theorem led to the publication of a remarkable discovery. He found that as long as the exceptional points are distributed in a carefully specified way on the x axis there can be infinitely many of them. The most important step in the proof of the new result was the precise description of the infinite set of exceptional points, and for this purpose Cantor realized he needed to provide a satisfactory way of analyzing the continuum of points along the x axis. Thus Cantor's study of trigonometric series gave rise to an important transition in his thought: he began to focus his attention on the relations among points in the continuum instead of on theorems about trigonometric series.

Cantor regarded it as axiomatic that for any point on a continuous line there corresponds a number, which is called a real number in order to distinguish it from the "imaginary" numbers that are multiples of $\sqrt{-1}$. Conversely, for every real number there corresponds exactly one point on a continuous line. Hence the problem of describing the continuum of points on a line is equivalent to the problem of defining the real numbers and exploring their properties. One of the major accomplishments of Cantor's paper of 1872 was a rigorous account of these properties.

The main obstacles to a theory of the real numbers are numbers such as π and $\sqrt{2}$, which are not rational numbers. (A rational number is a number that can be expressed as the quotient of two whole numbers. It was known in antiquity that $\sqrt{2}$, $\sqrt{3}$, $\sqrt{5}$ and many other roots are irrational.) Because the legitimacy of the rational numbers was not in question, Cantor followed an approach suggested by Karl Weierstrass, one of his former teachers at the University of Berlin. Cantor proposed that any irrational number could be represented by an infinite sequence of rational numbers. The number $\sqrt{2}$, for example, can be represented by the infinite sequence of rational numbers 1, 1.4, 1.41 and so on. In this way all irrational numbers can be understood as geometric points on a real-number line just as the rational numbers can.



SMOOTH, CONTINUOUS GRAPH whose height depends on the value of a corresponding point on the x axis can be approximated to any desired accuracy by a trigonometric series, that is, by a sum of sines and cosines. For example, a straight, horizontal line one unit above the x axis (*color*) can be approximated by superposing sine waves on one another (*gray curves*); the first two stages of the approximation are depicted (*black curves at top and middle*). The trigonometric series that approximates the graph is unique. Even if the graph is not continuous, however, it can often be approximated by a unique trigonometric series. For example, if the height of the graph is one unit everywhere except at the point where x is equal to 1/2, the trigonometric series that converges to the continuous line also converges to the broken line except at the point 1/2 (*bottom*). Cantor showed that a graph can be approximated by a unique trigonometric series even if the number of points where the graph is not continuous is an infinite number, provided the points of discontinuity are distributed along the x axis in a specific way. He was thereby led to consider the structural properties of abstract infinite sets and the infinitely many ways in which elements of infinite sets can be ordered with respect to one another.

In spite of the advantages of Cantor's approach, it troubled some mathematicians because it presumes the existence of sets or sequences of numbers having infinitely many elements. Philosophers and mathematicians had rejected the concept of completed infinities since the time of Aristotle, primarily because of the logical paradoxes they inevitably seemed to generate. Galileo, for example, pointed out that if completed infinite sets are admissible in mathematics, there must be as many even integers as there are even and odd integers together. Every even integer can be uniquely matched with the integer that is half its magnitude and so there is a one-to-one correspondence between each member of one set and each member of the other.

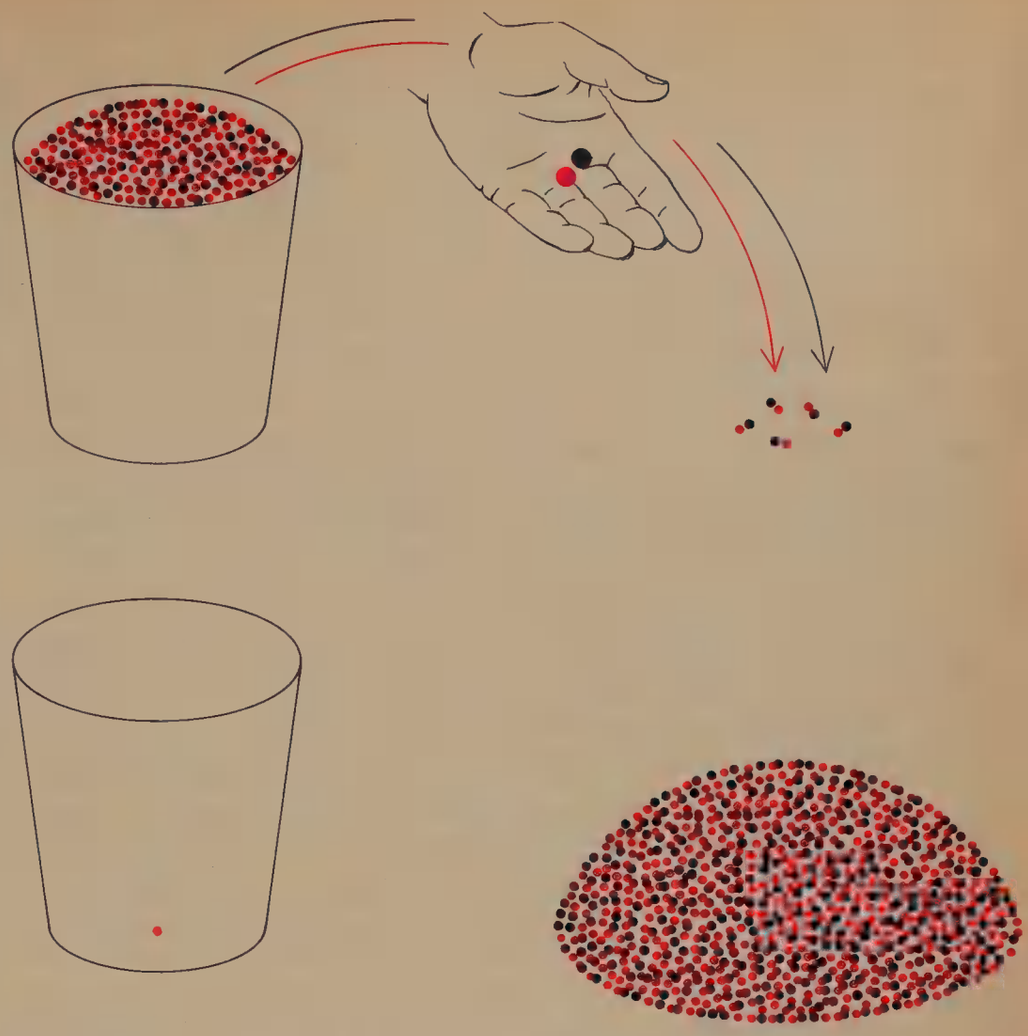
Theologians such as St. Thomas Aquinas, representing another voice of traditional opposition to the completed infinite, regarded the idea as a direct challenge to the unique and absolutely infinite nature of God.

In order to meet such objections mathematicians had customarily drawn a sharp distinction between the infinite regarded as a completed quantity and the infinite regarded as potential, represented by an indefinite sum or series of numbers that tends toward some limit. Only the potential infinite were they willing to countenance. In 1831 Carl Friedrich Gauss expressed his opposition to the use of completed infinities in terms Cantor once described as authoritative. In a letter to Heinrich Schumacher, Gauss wrote: "But concerning your proof I protest above all the use of an infinite quantity as a *completed* one, which in mathematics is never allowed. The infinite is only a *façon de parler,* in which one properly speaks of limits."

By speaking of limits it was possible to avoid the paradoxes associated with actual infinities. For example, by adding additional digits to the decimal expansion of π one can approximate the true value of π with increasing precision. Gauss insisted, however, that one must never assume all the terms of the decimal expansion could be given, so that the numerical value of π would be determined exactly. To do so would be to grasp an infinite number of terms in its entirety, or in other words an actually infinite set of numbers, an operation Gauss refused to allow.

Cantor was not alone in studying the properties of the continuum in rigorous detail. In 1872, the same year Cantor's paper appeared, the German mathematician Richard Dedekind also published an analysis of the continuum that was based on infinite sets. In his paper Dedekind articulated an idea that Cantor later made more precise: "The line is infinitely richer in point-individuals than is the domain . . . of rational numbers in number-individuals." The property can be put into perspective if one considers the distribution of the points on a line segment that correspond to rational numbers, or rational points. No matter how small the segment, there are infinitely many rational points on it. Thus the gist of Dedekind's remark was that in spite of the density of the rational points on a line segment there is still room to pack an infinite number of irrational points into the line. Irrational points such as $\sqrt{2}$ fall between the rational points, and so the set of rational points, although dense, is still riddled with holes and fails to be continuous.

Dedekind's statement is consistent with a correct understanding of the continuum, but it conceals a serious weakness. If anyone had asked Dedekind how much richer the infinite set of points in the continuum was than the infinite set of rational numbers, he could not have replied. Cantor's major contribution to the question was published in 1874 by August Leopold Crelle's *Journal für die reine und angewandte Mathematik,* also referred to as *Crelle's Journal* and perhaps the most respected mathematical periodical of its day.

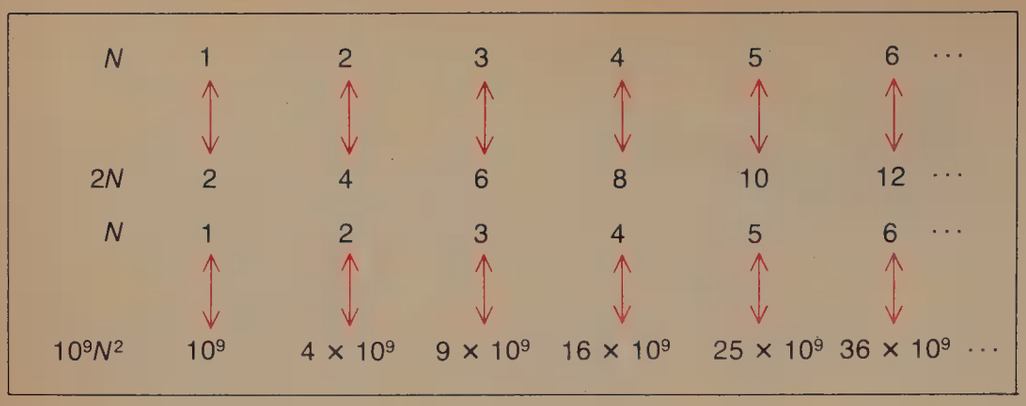


TWO SETS can be compared in size by matching the elements in the first set one for one with the elements in the second. For example, in order to determine whether or not there are more colored marbles than black marbles in a bucket the marbles can be removed from the bucket in pairs of one colored and one black marble. When no more pairings can be made, the marbles remaining in the bucket, if any, are the basis of the comparison. Cantor applied the same elementary principle in order to describe how the sizes of infinite sets can be compared.

In effect Cantor borrowed the paradox cited by Galileo and turned it into a means of comparing the size of infinite sets. He defined two sets as equivalent if a one-to-one correspondence can be established between the members of each set. Such a correspondence provides a natural way of comparing size. For example, imagine a bucket filled with black and colored marbles, and suppose one wants to compare the number of black marbles with the number of colored ones. The simplest way is to remove the marbles from the bucket in pairs of one black and one colored marble. If every marble can be paired with a marble of a different color, the two sets are equivalent. Failing that, the marbles remaining in the bucket are the basis of the comparison.

By applying the same principle of correspondence Cantor showed that the property Galileo regarded as a paradox was actually a natural property of infinite sets. The set of even numbers is equivalent to the set of all positive whole numbers, even and odd together, because the pairings between members of each set can go on forever without omitting any members from either set. Cantor was also able to exhibit an ingenious and sophisticated method whereby the set of all rational numbers could be paired with the whole numbers [*see bottom illustration on next page*]. Any set of numbers whose members can be matched one for one, or in effect counted by the set of positive whole numbers, Cantor called a denumerable set.

Given the density of the rational numbers on a line and the relative sparseness of the whole numbers, it may seem grossly counterintuitive that the two sets turn out to be the same size. Cantor's next move, however, was even more striking. He showed that no such one-to-one correspondence can exist between the set of whole numbers and the set of all points on a line, that is, the real numbers; in short, the real numbers form

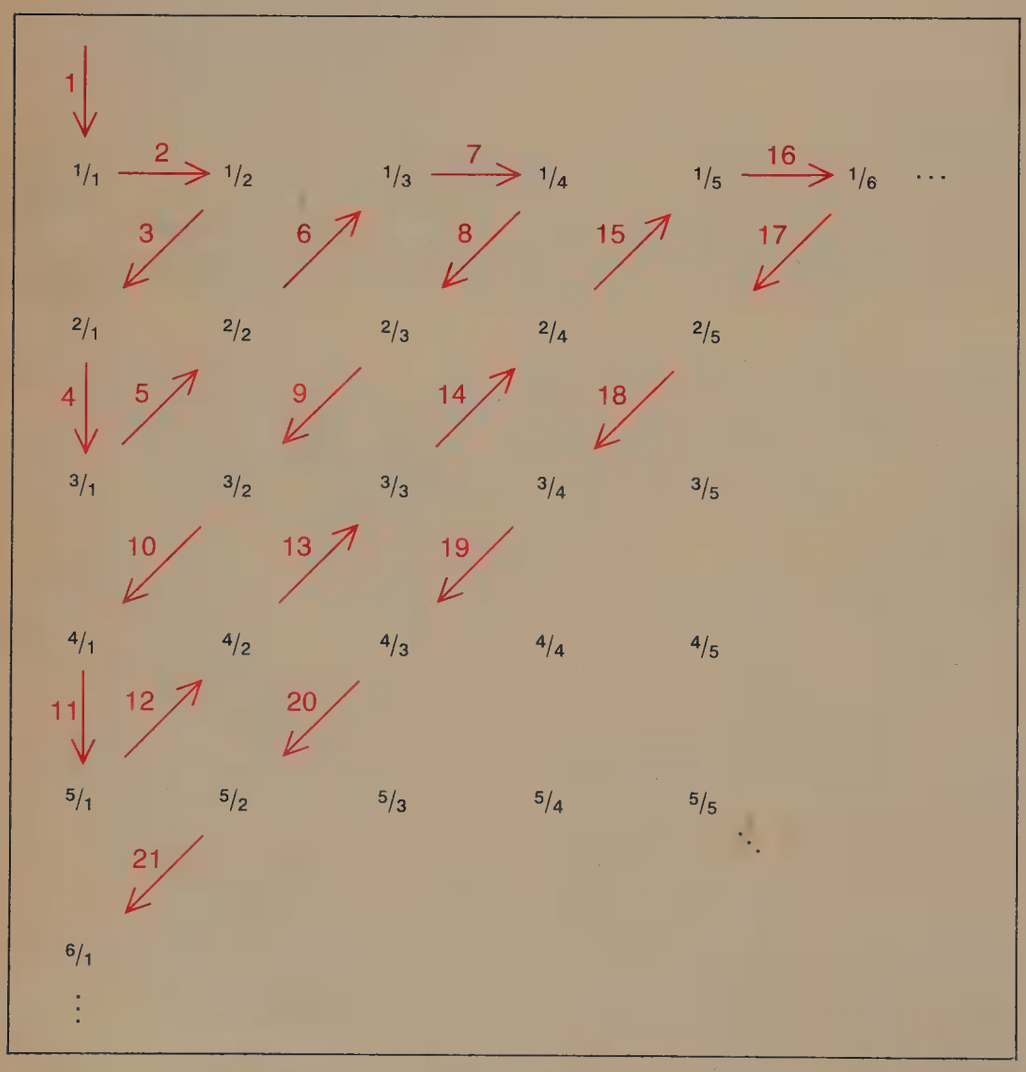


WHOLE NUMBERS can be paired one for one with the even numbers without exhausting either set of numbers. Hence although it may appear there must be more whole numbers than even numbers, the two infinite sets actually have the same number of elements. Many other infinite sets, such as the perfect squares multiplied by a billion, can also be compared one for one or, in effect, counted by the whole numbers. Such sets are called countable or denumerable sets.

a nondenumerable set. Cantor gave a somewhat complicated proof for this assertion in his paper of 1874; I shall instead present the main idea of a simplified and much more powerful version he gave in 1891.

Cantor began his proof by supposing there does exist a correspondence between the set of real numbers and the set of whole numbers. The argument then proceeds by showing that the supposition leads to a contradiction; it therefore follows that the original supposition must be false and that no such one-to-one correspondence is possible. The argument can be simplified by examining only the real numbers between 0 and 1; if this set of real numbers is greater than the set of whole numbers, the set of all the real numbers is greater than the set of whole numbers as well.

Suppose, therefore, that the real numbers between 0 and 1 can be paired one for one with the whole numbers. Setting up such a correspondence is equivalent to making a list of the real numbers, each one represented as an infinite decimal. It is then possible to define a new real number that has not been included in the list. Consider the first digit of the first decimal expansion in the list of real numbers. If the digit is a 1, write a 9 in the first decimal place of the defined number. If the first digit in the list is not a 1, write a 1 in the first place for the defined number. Continue constructing the new number to be defined by changing the second digit of the second decimal expansion in the list, the third digit in the third decimal expansion and so on. The newly constructed number must differ in at least one decimal place from every real number already in the list, but it will still represent some real number between 0 and 1. A number has therefore been constructed that is not in the list of real numbers, and so the supposition that all the real numbers can be listed leads to a contradiction.



INFINITE SET OF RATIONAL NUMBERS (that is, numbers that can be expressed as the quotient of two whole numbers) might seem much larger than the set of whole numbers. For example, between any two consecutive whole numbers such as 0 and 1 there are infinitely many rational numbers. Nevertheless, Cantor showed in 1874 how the rational numbers can be paired one for one with the whole numbers. Every possible rational number can be listed in a square array as shown in the illustration; each rational number can then be associated with a whole number by tracing the path shown in color. Thus the set of rational numbers is denumerable.

In August, 1874, Cantor married Vally Guttmann; the couple spent the summer in the Harz Mountains, where they also met with Dedekind. The period was an exceedingly fruitful one for Cantor's work. Earlier that year he had posed his next important problem to Dedekind in a letter: "Might it be possible to correspond a surface (a square, perhaps, including its boundaries) with a straight line (perhaps an interval with the inclusion of its endpoints) so that to each point of the surface one point of the line corresponds, and conversely?" Although Cantor thought the answer must be negative, he could offer no reason for his belief, and neither could Dedekind.

By 1877, however, Cantor was able to send Dedekind the startling news that, contrary to prevailing mathematical opinion, a one-to-one correspondence between lines and planes is not impossible. The proof is a matter of representing each point on a square by a pair of decimal coordinates. The decimal representations are then shuffled together in a strictly specified way to generate a single decimal expansion; this decimal is associated with a point on a line segment. The entire process can also be reversed [see illustration on page 129]. Cantor himself was so unprepared for the result that it prompted him to exclaim, "I see it but I don't believe it!"

Cantor immediately prepared a manuscript that described his new discovery, and he sent it, as he had his paper of 1874, to *Crelle's Journal*. Although the

result was of major significance, the paper was also the first occasion for open hostilities between Cantor and his former teacher Kronecker. As an editor of the journal, Kronecker was in a position to block the publication of any article, and by 1877 he was so appalled at the direction of Cantor's work that he did just that. Although Cantor had submitted his manuscript on July 12, no steps were taken to prepare it for publication and it did not appear in the volume for 1877. Cantor, suspecting Kronecker's intervention, wrote a bitter letter to Dedekind complaining about the treatment of his paper and mentioning the possibility of withdrawing it from the journal. Dedekind, recalling his own experience in such matters, persuaded Cantor to wait. As it turned out, Dedekind was right. The article appeared in the volume for 1878, but Cantor was so offended by the incident that he refused to publish in *Crelle's Journal* again.

The controversy that had arisen between Cantor and Kronecker was clouded with personal animosity, but at its roots were sharply divergent views on the philosophy of mathematics. Such views are still reflected in the debate between constructivist and formalist mathematicians. Kronecker, a forerunner of the constructivists, is well known for his quip that captures the essence of his position: "God made the integers; all else is the work of man." In this spirit he advocated constructing all mathematics from the integers and from finite arithmetic combinations of them. In the early 1870's he began to reject any of the limiting processes in the traditional calculus, and he opposed all attempts to define mathematical objects in terms of limits. Thus even the irrational numbers, which had been accepted by mathematicians for centuries, were to be read out of mathematics unless some way could be found to construct them, as the rational numbers are constructed, out of the integers.

Cantor, who had written two major papers under Kronecker in his student days at the University of Berlin, was well aware of Kronecker's extreme position and was sensitive to its advantages. It guaranteed the maximum certainty and correctness to a mathematical proof, and it could be applied as a corrective to unfettered speculation in mathematics. Cantor nonetheless argued that to accept Kronecker's position would be to rule out many of the most promising developments in mathematics; moreover, it could burden innovative mathematical research with narrow and ultimately barren methodological scruples.

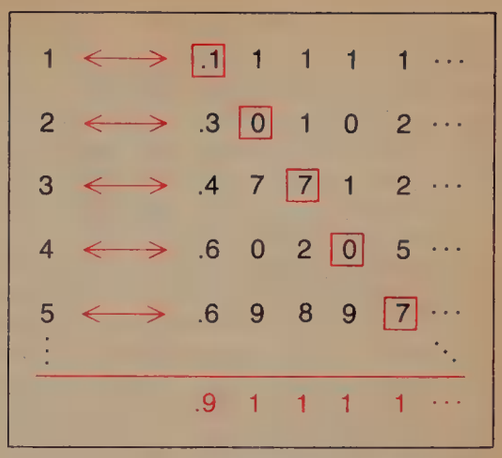
The definition of the irrational numbers given in Cantor's 1872 paper was tantamount to accepting the existence of completed infinite sets. Cantor embraced a formalist position on the existence of the irrationals and he argued that the only grounds on which their legitimacy was to be judged in mathematics were their formal and internal consistency. "In introducing new numbers," he once wrote, "mathematics is only obliged to give definitions of them, by which... they can definitely be distinguished from one another. As soon as a number satisfies all these conditions it can and must be regarded as existent and real in mathematics."

This point of view on the irrational numbers was crucial for Cantor's justification of his introduction of transfinite numbers. In his 1872 paper he had defined sets of exceptional points by introducing the concept of a limit point. The irrational number $\sqrt{2}$, for example, is a limit point of the sequence 1, 1.4, 1.41 and so on. More generally, a point is a limit point of an infinite set if there are an infinite number of members in the set that lie arbitrarily close to the point.

Given any infinite set P, Cantor defined the derived set P^1 as the set of all the limit points of P. Similarly, if P^1 were also an infinite set, its derived set P^2 was defined as the set of all the limit points of P^1. Cantor showed that the relation of being a subset determines a natural ordering for the sets: it turns out that every member of P^2 is also a member of P^1 and so P^2 is a subset of P^1; similarly, P^3 is a subset of P^2 and so on.

It can turn out that for some finite whole number n the derived set P^n is a finite set; if this condition holds, the infinite set of points P that gives rise to the set P^n is precisely the set of exceptional points needed to prove the general version of Cantor's theorem about the unique representation of functions by trigonometric series. On the other hand, it can also happen that no derived set in the sequence $P^1, P^2, P^3, \ldots$ is a finite set. Cantor argued that in this case it still makes sense to consider the set of points that are common to all the derived sets $P^1, P^2, P^3, \ldots, P^n, \ldots$. He designated the set of points common to all these derived sets P^∞, and in 1880 he began to refer to the superscript ∞ as a transfinite symbol. Moreover, if P^∞ were an infinite set of points, one could then form its derived set $P^{\infty+1}$, which could in turn lead to an entire sequence of derived sets, $P^{\infty+2}$ and so on.

Cantor might have added that the superscripts ∞, $\infty + 1$, $\infty + 2$ and so on actually constitute a new kind of number, but at first he did not do so. In 1872 he had been careful to discuss the irrational numbers only in terms of sequences of rational numbers; similarly, he initially regarded the symbols ∞, $\infty + 1$, $\infty + 2$ and so on only as tags for identifying sets. In 1883, however, he abandoned his reticence and presented them as transfinite numbers, an autonomous and systematic extension of the natural numbers.



SET OF REAL NUMBERS, represented by the continuum of points on a line, is not denumerable. If it were, the real numbers between, say, 0 and 1 could be paired one for one with the whole numbers. Each real number in the list can be represented by an infinite decimal expansion. (Infinite decimals such as .5000... are to be represented by an equivalent infinite decimal, such as .4999....) No matter how such infinite decimals are listed, however, a new decimal can be constructed that defines a real number not included in the list: write 9 in the first decimal place of the constructed number if the first digit in the decimal expansion of the first real number in the list is a 1; otherwise write 1 in the first decimal place of the constructed number. Continue by changing the second decimal digit of the second real number, the third decimal digit of the third real number and so on. The constructed decimal expansion represents a real number between 0 and 1, but it must differ in at least one decimal place from every real number in the list. Hence the supposition that the real numbers can be paired with the whole numbers leads to a contradiction and so must be abandoned. The main idea employed in the proof is called the diagonalization method.

The immediate reason for their introduction, Cantor maintained, was that they were necessary for further progress in set theory and for the study of the real numbers. Nevertheless, in order to respond to critics such as Kronecker, Cantor argued from his formalist philosophical position: Once the self-consistency of the transfinite numbers was recognized they could not be refused a place alongside the other accepted but once disputed members of the mathematical family, such as the irrational numbers. By formulating a theory of the infinite that was able to avoid the well-known mathematical paradoxes, Cantor believed he had removed the only rational objection mathematicians could raise for refusing to consider the completed infinite in their work.

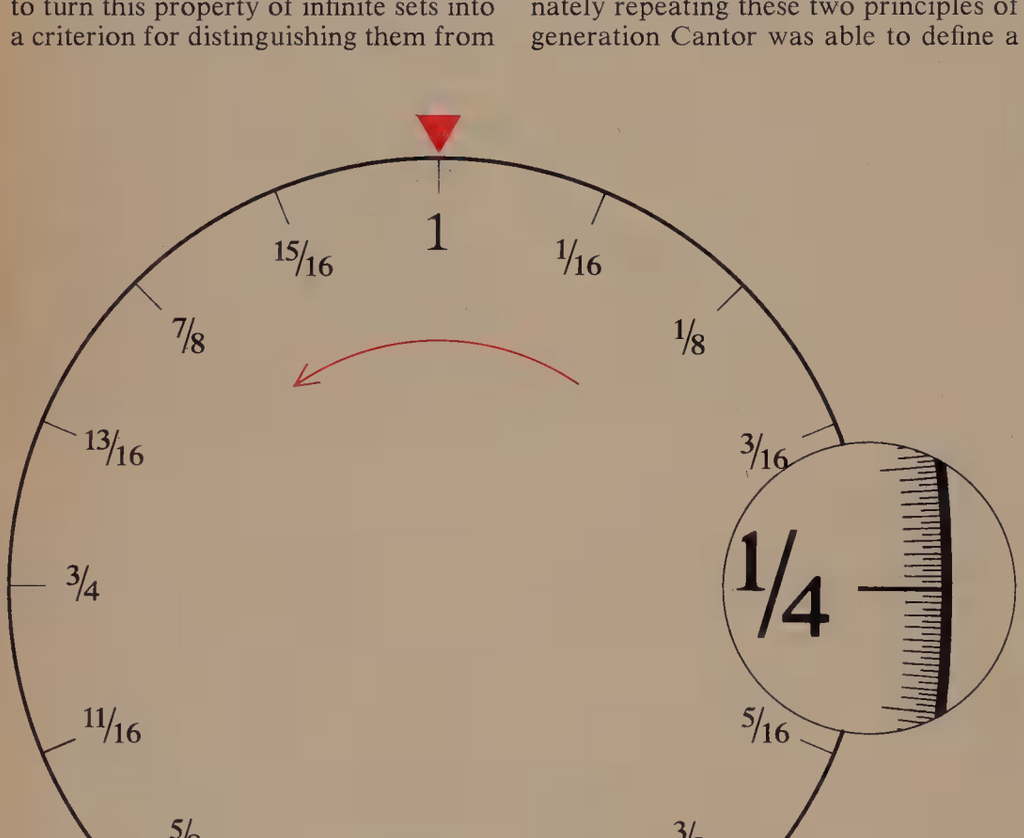
The transfinite numbers ultimately introduced by Cantor are widely known by the notation he adopted for them in later years, namely the first letter of the Hebrew alphabet ℵ (aleph). The alephs designate the cardinality, or number of elements, in an infinite set, and so the

equivalences among infinite sets that Cantor demonstrated in the 1870's are frequently stated in terms of the transfinite cardinal numbers, the alephs. It is therefore of considerable historical interest that the first transfinite numbers to be introduced were not cardinal numbers but ordinal numbers.

An ordinal number is defined by its order or position in a list. The ordinal number associated with a finite set corresponds to the cardinal number of the set. For example, any set having five elements (that is, any set whose cardinal number is five) can in a certain sense be thought of as the immediate successor to any set having four elements. In other words, the ordinal number of the set is also five; it is the fifth set in a list of sets. The ordinal number of an infinite set, however, must be distinguished from its cardinal number. Cantor showed it is possible to construct an infinite number of infinite sets having different ordinal numbers but the same cardinal number. Indeed, Cantor later found it is possible to turn this property of infinite sets into a criterion for distinguishing them from finite sets. A set is finite only if its cardinal number and its ordinal number are the same.

Cantor pointed out that the ordinal number of a sequence of finite sets of increasing size 1, 2, 3 and so on is based on the repeated addition of units. There is no greatest ordinal number associated with the sequence of finite sets, but just as it is possible to define π as the limit of a sequence of rational numbers without supposing π is itself a rational number, so, Cantor believed, it is proper to define a new, transfinite ordinal number ω as the first number following the entire sequence of ordinary ordinal numbers 1, 2, 3 and so on. Once ω is defined it is possible through the repeated addition of units to generate additional transfinite ordinal numbers, $\omega + 1$, $\omega + 2$, $\omega + 3$ and so on. Because there is no largest element for this sequence either, one could imagine another ordinal number $\omega + \omega$, or 2ω, as the first ordinal following the sequence $\omega + 1$, $\omega + 2$, $\omega + 3$ and so on. By alternately repeating these two principles of generation Cantor was able to define a hierarchy of progressively greater transfinite ordinal numbers [*see illustration on page 130*].

How does one distinguish between, say, the ordinal number ω and the ordinal number $\omega + 1$? The difference is determined by the order of the elements within the sets to which ω and $\omega + 1$ correspond. For example, the set of natural numbers in its familiar sequence $(1, 2, 3, \ldots)$ has the ordinal number ω, which represents the entire sequence of natural numbers in its usual order. The set of all the natural numbers in the rearranged sequence $(2, 3, 4, \ldots, 1)$, however, or the set of all natural numbers in the sequence $(10, 30, 40, \ldots, 20)$ has the ordinal number $\omega + 1$. In other words, the distinction depends on the consecutive order of the elements in the sequence and on the placement of the infinitely long gap, represented by the ellipsis; if one number is displaced to the end of the sequence, the ordinal number of the new sequence is $\omega + 1$. The sequence $(2, 4, 6, \ldots, 1, 3, 5, \ldots)$ has two infinite gaps, and its ordinal number is $\omega + \omega$, or 2ω. Note that all the sets have the same number of elements, that is, their elements can always be placed in a one-to-one correspondence with each other and with the positive whole numbers. Hence their cardinal numbers are the same even though their ordinal numbers are quite different.



PROBABILITY OF RANDOMLY SELECTING a point that represents a rational number from the continuum of real numbers illustrates how the set of rational numbers compares in size with the set of real numbers. The probability is the ratio of the number of rational-valued, or rational, points to the total number of points along some interval. Here the interval between 0 and 1 is represented by the circumference of a wheel of fortune. (On the wheel 0 and 1 are identified.) It is assumed that by spinning the wheel and allowing it to stop any point on the wheel can be selected at random. The points representing the rational numbers are infinitely dense, in the sense that along any arc between two rational points on the circumference, no matter how short, there must be an infinite number of rational points. A few such points are marked. Nevertheless, the set of all the points along the circumference is infinitely larger than the set of rational points; the probability that the wheel stops at a rational point is zero. More precisely, the probability is smaller than any positive quantity that can be specified in advance.

Once the transfinite ordinal numbers were defined Cantor proceeded to describe their arithmetic properties. An important distinction between the transfinite and the ordinary numbers had to be made with respect to the commutative property of addition and multiplication. For two ordinary numbers A and B the commutative property expresses the fact that $A + B$ is equal to $B + A$ and $A \times B$ is equal to $B \times A$. When addition or multiplication are defined to include the transfinite numbers, however, the commutative property cannot be assured. For example, $\omega + 2$, which represents the sequence $(1, 2, 3, \ldots, 1, 2)$, is not equal to $2 + \omega$, which represents the sequence $(1, 2, 1, 2, 3, \ldots)$.

Although the difference between ordinal number and cardinal number is indistinct for finite sets, it helped to explain how the application of the concept of number to an infinite set could lead to confusion and paradox. Because the concepts of ordinal and cardinal number are fundamentally distinct for infinite sets, any argument that discusses the number associated with an infinite set without making the distinction is subject to ambiguity. Thus it is illegitimate to extend the seemingly unambiguous properties of finite sets to infinite sets, as Galileo and others had done.

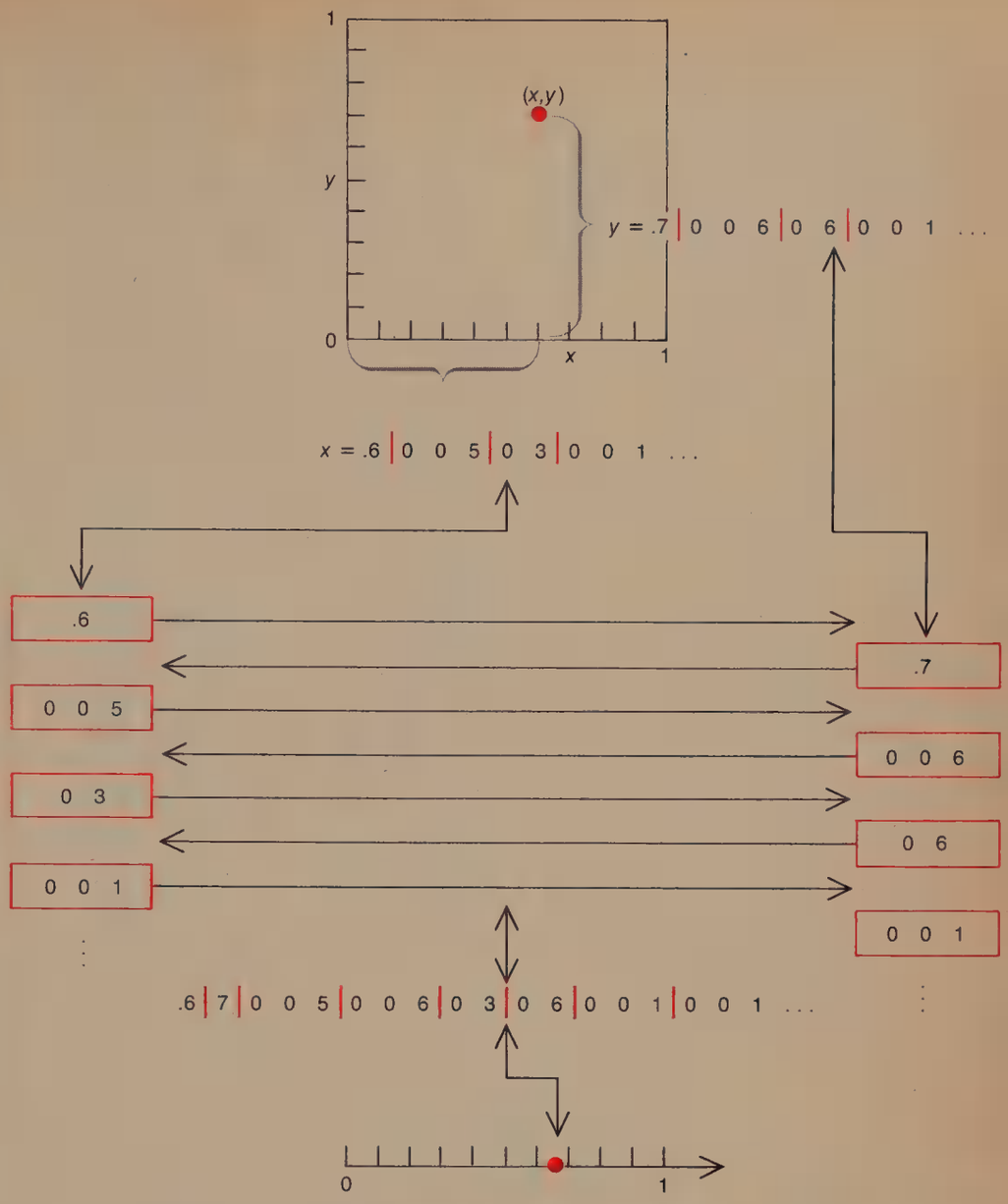
In spite of the results achieved by Cantor in the 1880's there remained a serious lacuna. The question of the cardi-

nal number (or power, in Cantor's early terminology) of the continuum of real numbers was still unanswered. Remember that in his paper of 1883 Cantor had defined the sequence of transfinite ordinal numbers according to two principles of generation. In order to introduce natural divisions in the sequence he had added a third principle. Consider the set of all finite whole numbers, which Cantor called the first number class. Its power or cardinal number is greater than the power associated with any element in the set. Similarly, he noted, one could also consider the set of all the transfinite ordinal numbers corresponding to denumerably infinite sets, or in other words sets whose power is the same as the power of the set of all whole numbers. Cantor called this set of transfinite ordinal numbers the second number class. It turns out that the power of the second number class is greater than the power associated with any of the transfinite ordinal numbers that make up the set. In short, the second number class is a nondenumerable set. Although Cantor was never able to prove it, he was convinced that the power of the second number class is equivalent to the power of the continuum of real numbers.

The conjecture has come to be known as Cantor's continuum hypothesis, and it has never been proved. In 1963 Paul J. Cohen of Stanford University, building on the work of Kurt Gödel of the Institute for Advanced Study, showed that although the continuum hypothesis is consistent with the axioms of a standard version of set theory, it is also independent of them. In fact, the continuum hypothesis plays the role for set theory that Euclid's parallel postulate plays for geometry. Different versions of set theory can be constructed by assuming the truth or falsity of the continuum hypothesis, just as Euclidean or non-Euclidean geometry can be constructed by assuming the truth or falsity of the parallel postulate [see "Non-Cantorian Set Theory," by Paul J. Cohen and Reuben Hersh; SCIENTIFIC AMERICAN, December, 1967].

Cantor's unsuccessful efforts to prove the continuum hypothesis caused him considerable stress and anxiety. Early in 1884 he thought he had found a proof, but a few days later he reversed himself completely and thought he could disprove the hypothesis. Finally he realized he had made no progress at all. Throughout the period he had to endure mounting opposition and threats from Kronecker, who said he was preparing an article that would show "the results of modern function theory and set theory are of no real significance."

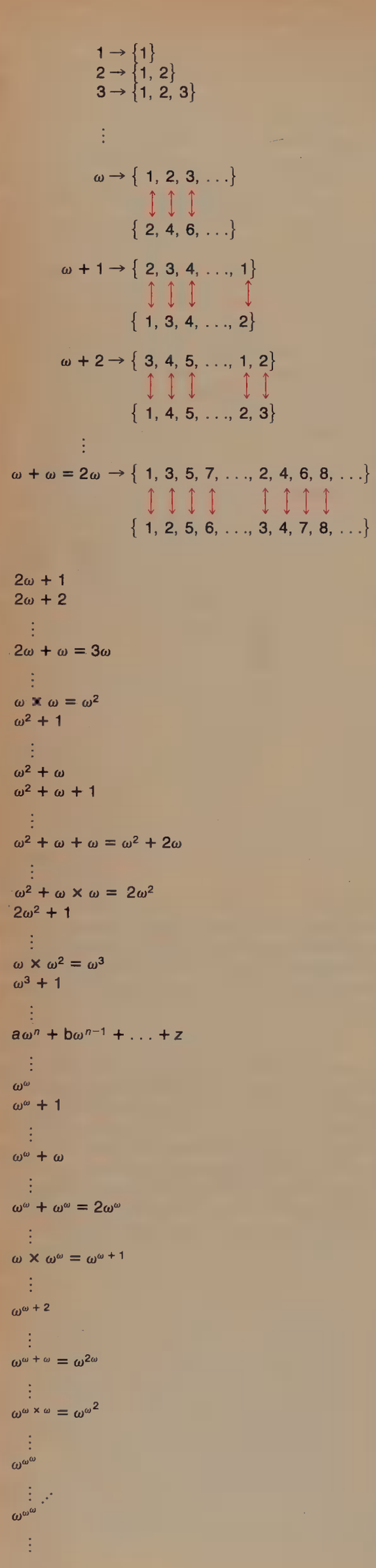
Soon thereafter, in May, 1884, Cantor had the first of his serious nervous breakdowns. Although frustration over his lack of progress on the continuum hypothesis or stress from Kronecker's attack may have helped to trigger the breakdown, it now seems clear that such events had little to do with its underlying cause. The illness took over with startling speed and lasted for somewhat more than a month. At the time only the manic phase of manic-depressive psychosis was recognized as a symptom; when Cantor "recovered" at the end of June, 1884, and entered the depressive phase, he complained that he lacked the energy and interest to return to rigorous mathematical thinking. He was content to take care of trifling administrative matters at the university and felt capable of little more.



POINTS IN THE PLANE can be matched one for one with the points on the line. Each point in the plane is represented by a pair of infinite decimals, and the infinite decimals are broken up into groups. Every digit except 0 in the decimal expansions begins a new group. The groups are then combined by alternating them into a single infinite decimal, which represents a point on the line. The entire procedure can also be reversed. A similar proof shows that the number of points in any finite-dimensional space is equivalent to the number of points on the line.

Although Cantor eventually returned to mathematics, he also became increasingly absorbed in other interests. He undertook a study of English history and literature and became engrossed in a scholarly diversion that was taken with remarkable seriousness by many people at the time: the supposition that Francis Bacon was the author of Shakespeare's plays. Cantor also tried his hand without success at teaching philosophy instead of mathematics, and he began to correspond with several theologians who had taken an interest in the philosophical implications of his theories about the infinite. The correspondence was of special significance to Cantor because he was convinced that the transfinite numbers had come to him as a message from God. He was anxious that his views be carefully studied by theologians so that his mathematical concept of the infinite could be reconciled with church doctrine.

More important, Cantor was instrumental in creating a professional society for the promotion of mathematics in Germany, the Deutsche Mathematiker-Vereinigung. He believed his own career had been greatly damaged by the premature and prejudiced rejection of his



work by the mathematics establishment, and he hoped the independent organization would give younger mathematicians encouragement and a fair hearing for new and perhaps radical ideas.

There was one last element of transfinite set theory with which Cantor had to come to terms, namely the nature and status of the transfinite cardinal numbers. The evolution of his thinking on the subject is curious, because the transfinite cardinal numbers were the last part of his theory to be given either rigorous definition or a special symbol. Indeed, it is difficult in the clarity of hindsight to reconstruct the obscurity within which Cantor must have been groping, and I have discussed his work up to now as if he had already recognized that the power of an infinite set could be understood as a cardinal number. In fact, although Cantor already understood it is the power of a set that establishes its equivalence (or lack thereof) with any other set, he initially avoided any suggestion that the power of an infinite set could be interpreted as a number.

He began to equate the two concepts in September, 1883; still, however, no symbol was provided for distinguishing one transfinite cardinal number from another. Since he had already adopted the symbol ω to designate the least transfinite ordinal number, it is clear that the ordinal numbers were much more important than the cardinals for the early conceptual development of Cantorian set theory. When Cantor finally introduced a symbol for the first transfinite cardinal number, it was borrowed from the symbols already in service for the transfinite ordinals; the first transfinite cardinal was written $\overset{*}{\omega}$.

Cantor did not settle on the alephs as symbols until 1893. By then the Italian mathematician Giulio Vivanti was preparing a general account of set theory, and Cantor realized it was time to adopt a standard notation. He chose the alephs for the transfinite cardinals because he thought the familiar Greek and Roman alphabets were already too widely employed in mathematics for other purposes. His new numbers deserved something distinct and unique. Thus he chose the letter $\aleph$, which was readily available in the type fonts of German printers. The choice was particularly clever, as Cantor was pleased to admit, because the Hebrew aleph was also a symbol for the number 1. Since the transfinite cardinal numbers were themselves infinite unities; the aleph could be taken to represent a new beginning for mathematics. Cantor designated the cardinal number of the first number class $\aleph_0$ (aleph-null), the number he had previously called $\overset{*}{\omega}$; the cardinal number of the second number class was designated $\aleph_1$ (aleph-one).

Cantor made his last major contributions to set theory in a pair of articles published in 1895 and 1897. He had proved in a paper delivered before the first meeting of the Deutsche Mathematiker-Vereinigung in 1891 that the cardinal number of any set is always smaller than the cardinal number of the set of all its subsets. (A version of the proof is given in the illustration on the opposite page.) A few years later he derived a corollary to this result: the cardinal number of the continuum is equal to a cardinal number Cantor designated $2^{\aleph_0}$. He hoped this result would soon lead to a solution of the continuum hypothesis, because the hypothesis could now be stated in clear algebraic form: $2^{\aleph_0} = \aleph_1$.

The arguments in Cantor's proof about the cardinal number of the set of subsets, however, led to far different conclusions. The most important of them was reached by Bertrand Russell in 1903; Russell showed that a paradox can be derived in set theory by considering all sets that do not include themselves as members. Russell's paradox suggested there was something fundamentally wrong with Cantor's definition of a set, and the consequences of this realization have become an important problem in 20th-century mathematical logic. Nevertheless, such developments have not overturned any of Cantor's major results in transfinite arithmetic.

Unfortunately by 1903 Cantor was suffering attacks of manic depression with increasing frequency, and no evidence has been found to suggest he ever became aware of Russell's result. In fact, illness prompted Cantor to apply for a leave of absence from the University of Halle during the fall term of 1899, and the request was granted. In

TRANSFINITE ORDINAL NUMBERS are identified by their order or position in a list. The list is generated according to two principles. First, each new transfinite ordinal is derived from the immediately preceding one by adding one unit, just as if one were "counting" beyond the transfinite ordinal ω that is associated with the set of whole numbers arranged in their natural order. Second, whenever a succession of transfinite ordinals exists for which there is no largest number, a new transfinite ordinal is defined as the next number larger than all the others. Such new numbers are listed immediately after a vertical mark of ellipsis; for example, 2ω is the next transfinite ordinal number larger than all the numbers ω, $\omega + 1$, $\omega + 2$ and so on. It is the placement of the infinite gaps (*horizontal marks of ellipsis*) within the sets to which the ordinal numbers correspond that distinguishes the transfinite ordinals from one another. Thus in the diagram two examples of sets are given that correspond to the ordinal numbers ω, $\omega + 1$, $\omega + 2$ and 2ω. Every infinite set represented by an ordinal number in the list, however, has the same cardinal number, namely $\aleph_0$, or in other words every set includes the same number of elements.

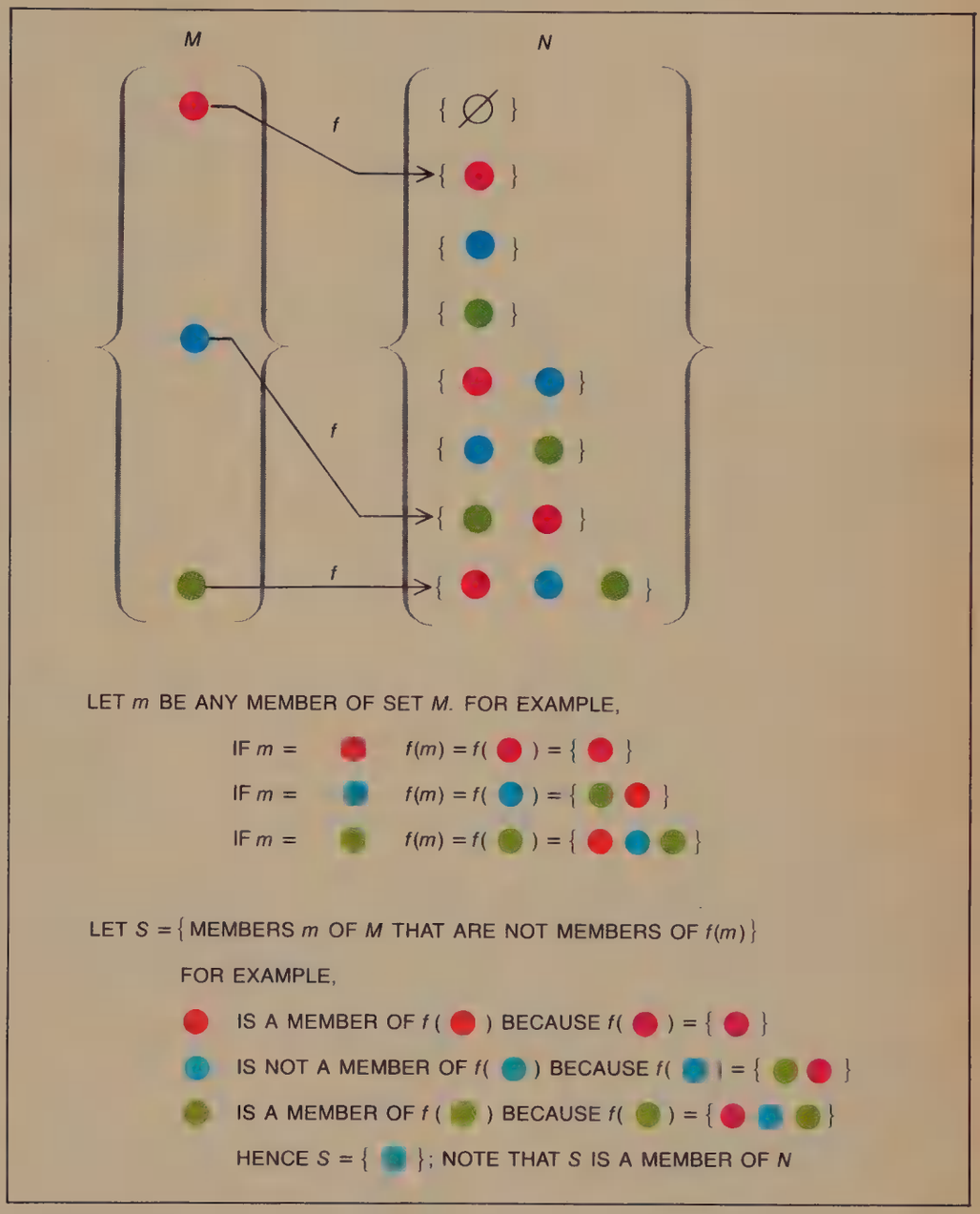
November of the same year he notified the German Ministry of Culture that he wanted to give up his professorship completely. As long as his salary was not reduced he would be content with a modest position in a library. Among his qualifications he emphasized his publications on the Shakespeare question, and he concluded with the extraordinary request that the Ministry issue its reply within the next two days. If he were offered no alternative to teaching, he wrote, then as a person born in Russia he would seek to enter the service of the Russian diplomatic corps.

Nothing seems to have come of Cantor's request, and he did not enter the service of Emperor Nicholas II. Nevertheless, the entire episode fits the earlier pattern of his behavior in 1884, when he had seriously considered giving up mathematics for philosophy after his first major breakdown. Similarly, he was hospitalized for manic depression late in 1899, again in the winter terms of 1902 and 1903 and thereafter for increasingly long and frequent periods. He died of heart failure on January 6, 1918, in the Halle Nervenklinik.

There is an important connection between Cantor's mental illness and his mathematics. Certain documents suggest that on occasion the illness afforded him periodic respite from daily affairs, during which he was still able to pursue his mathematical ideas in the solitude of the hospital or quietly at home. The illness may also have supported his belief that the transfinite numbers had been communicated to him from God. Following a long period of hospitalization in 1908, he wrote to a friend in Göttingen, the British mathematician Grace Chisholm Young. As he described it, his manic depression took on a strikingly generative quality: "A peculiar fate, which thank goodness has in no way broken me, but in fact has made me stronger inwardly, happier and more expectantly joyful than I have been in a couple of years, has kept me far from home—I can also say far from the world.... In my lengthy isolation neither mathematics nor in particular the theory of transfinite numbers has slept or lain fallow in me."

Elsewhere Cantor described his conviction about the truth of his theory in quasi-religious terms: "My theory stands as firm as a rock; every arrow directed against it will return quickly to its archer. How do I know this? Because I have studied it from all sides for many years, because I have examined all objections that have ever been made against the infinite numbers and above all because I have followed its roots, so to speak, to the first infallible cause of all created things."

Later generations might dismiss the philosophy, look askance at the abundant references to St. Thomas or to the church fathers, overlook the metaphysical pronouncements and miss entirely the deeply religious roots of Cantor's later faith in the absolute truth of his theory. But all these commitments contributed to his resolve not to abandon the transfinite numbers. Opposition seems to have strengthened his determination. His forbearance, as much as anything else he might have contributed, ensured that set theory would survive the early years of doubt and denunciation to flourish eventually as a vigorous, revolutionary force in the scientific thought of the 20th century.



INFINITE SEQUENCE OF SETS, each set larger than the one before it in the sequence, can be constructed by considering the set of all the subsets of any given set. An ingenious variation of Cantor's method of diagonalization can be employed to show that, by assuming a one-to-one correspondence *f* between any set *M* and the set of all its subsets *N*, it is possible to construct a subset *S* that is not included in the one-to-one correspondence, whatever *f* may be. In order to understand the construction consider the finite set *M* that includes a red, a blue and a green disk. The set has eight subsets (including the empty set Φ, which contains no elements). Let *S* be defined as the set of all the elements *m* of *M* that are not members of the subset *f*(*m*) to which *m* corresponds. For the example in the illustration *S* includes only the blue disk. Because *S* is a subset of *M* and because *f* is assumed to be a one-to-one correspondence, there must be some element *a* of *M* to which *S* corresponds, or in other words for which *f*(*a*) is identical with *S*. Either *a* is a member of *S* or it is not. If *a* is a member of *S*, it must also be a member of *f*(*a*), since *f*(*a*) is equal to *S*; on the other hand, if *a* is a member of *S*, it cannot be a member of *f*(*a*) because of the way *S* is defined. Hence *a* cannot be a member of *S*. Again, however, by the definition of *S*, if *a* is not a member of *S*, *a* must be a member of *f*(*a*), and since *f*(*a*) is equal to *S*, *a* must also be a member of *S*. No matter what the status of *a* is, therefore, the supposition that the set *M* can be paired one for one with the set of all its subsets leads to a contradiction, and so the supposition must be abandoned. In the same way it is proved that even if a set is infinite, the set of all its subsets is larger than the original set. An infinite sequence of progressively larger infinite sets can be constructed by forming the set *N* of all the subsets of any infinite set *M*, then forming the set *P* of all the subsets of *N* and so on. There is no largest set in the sequence.

The Slide Fastener

Otherwise known as the zipper, it took hold, so to speak, some 60 years ago. It is the creation of a host of inventors going back to Elias Howe of the sewing machine, and the end is not in sight

by Lewis Weiner

The slide fastener, perhaps more widely known as the zipper, appears on objects ranging from suitcases to trousers and dresses, and it would be unusual for anyone in a Western society to get through the day without encountering one. Yet this common and usually reliable mechanism goes remarkably uncelebrated. Out of curiosity I looked in seven leading encyclopedias, five published in the U.S., one in Britain and one in Germany. Some of them have in their latest edition an entry of a few lines on the subject; none has what could be called a full-scale article. The origin and development of the slide fastener are matters few people know about, even in the companies that make it. Having been associated with the industry for many years, partly as an inventor of machinery for making zippers and partly as a consulting engineer for manufacturers, I am moved to put forward an account of this simple and ubiquitous device.

In 1943 Frank B. Jewett, chairman of the Bell Telephone Laboratories and president of the National Academy of Sciences, gave an address at New York University titled "The Promise of Technology." Remarking that "the real creative ideas" originate with individuals and that no one can foretell what the ideas will be or where they may arise, he said: "As an example of how impossible it is, even with simple things, to forecast the future, I have often thought of how infinitesimally small would have been the chance of any man or group of men, except the one who actually had the idea, planning to invent the common zipper."

The one who actually had the idea was Elias Howe, who is much better known for his contribution to the invention of the sewing machine. In 1851 he received a U.S. patent for an automatic, continuous clothing closure. Describing it in the patent, he wrote: "My invention consists of a series of clasps united by a connecting cord running or sliding upon ribs." Howe's device anticipated many of the features of the modern slide fastener, but for reasons that are not clear he never put it on the market.

Hence the man who usually gets credit for inventing the zipper is Whitcomb L. Judson, who in 1893 received two U.S. patents for an automatic closing device actuated by a slide mechanism. Whether Judson had a limited vision of the applications that might be found for his fastener or simply wanted to state a specific application for the purpose of obtaining the patent is not known, but his title for the first patent was "Clasp Locker or Unlocker for Shoes." The fasteners differed from the modern zipper mainly in that their movement was perpendicular to the opening that was to be closed, like the buckles on the galoshes made before zippers were put on them. Judson's scheme consisted of a series of separate clasps, each of which consisted in turn of two interlocking parts, one attached to each side of the opening that was to be closed. They could be fastened by hand, but Judson also provided for a sliding device to close and open them in sequence.

In 1894 Judson and Lewis A. Walker organized the Universal Fastener Company to exploit the Judson patents. Walker was a lawyer who had a flair for corporate organization. The company began promoting one form of fastener under the name Universal in 1896. Sales were poor, partly because the fasteners had a tendency to pop open under stress and partly because they were rather sharp and sometimes tore fabrics that came in contact with them.

In 1904 the company was reorganized as the Automatic Hook and Eye Company and offered an improved fastener under the name C-Curity. (An advertising slogan was "A Pull and It's Done.") The company had the advantage of a machine Judson had patented in 1902 to make the fasteners, which previously had been made by hand. It had the further advantage in 1906 of employing Gideon Sundback, an engineer who over a period of years greatly improved the fastener.

Simultaneously other inventors were also improving it. In 1911 Catharina Kuhn-Moos and her partner Henri Foster received a Swiss patent showing a device closely resembling a modern metallic zipper in that it had no hooks. It appears these inventors were too far ahead of their time; their fastener was not a commercial success. In 1913 P. A. Aronson, the superintendent of Automatic Hook and Eye, was awarded a patent that shows the principle of the modern separating zipper: a fastener that opens at both ends.

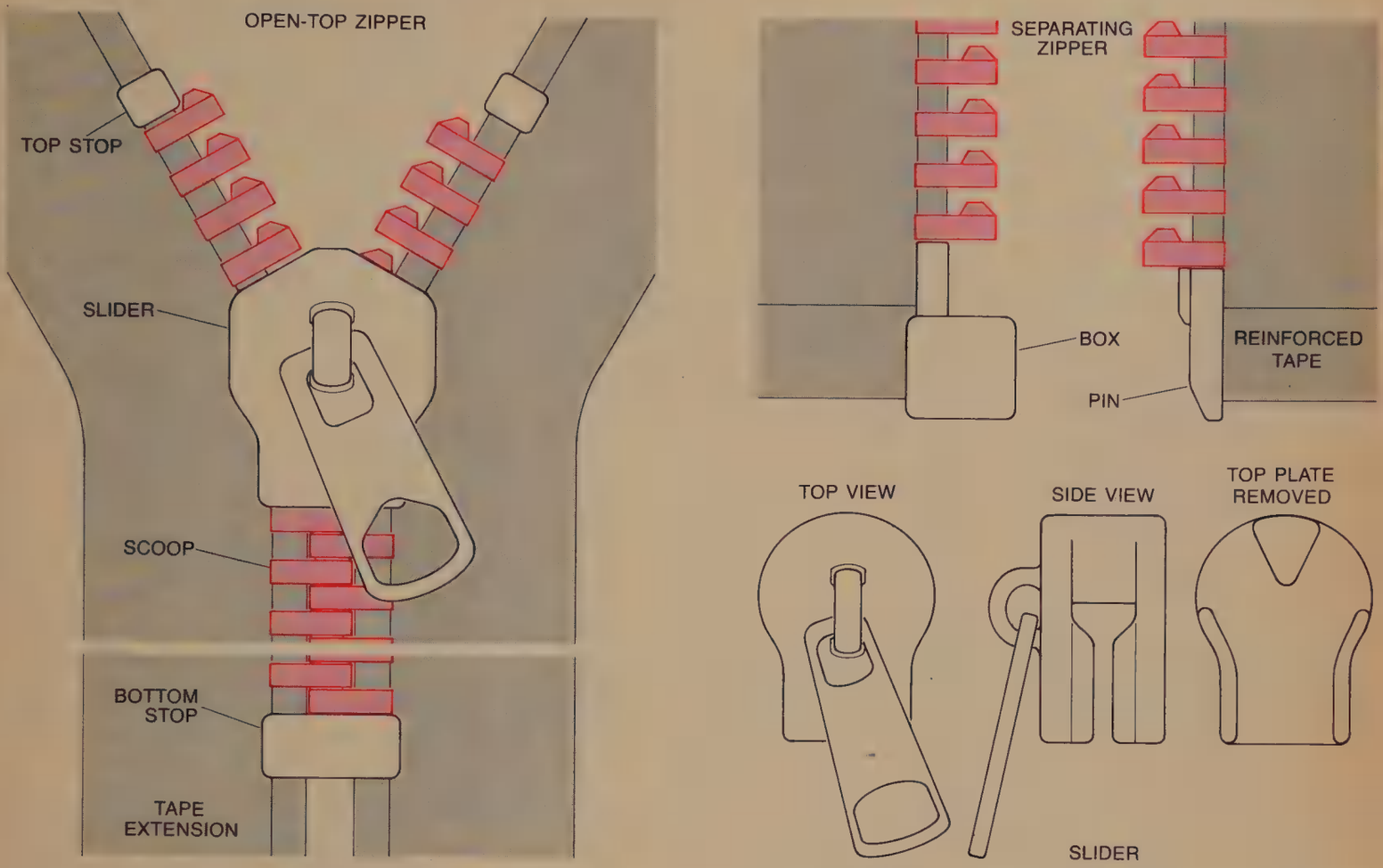
Sales continued to be poor, however, and the company verged on bankruptcy. It survived by taking on work other than fasteners. Then in 1917 Sundback received a patent on a metallic fastener that had all the features of the modern metallic zipper. In an interview he said: "The great need was to eliminate the hooks." The invention was so important to the company that it changed its name to the Hookless Fastener Company. Walker raised more capital and reincorporated the business with himself as president, a position he held until his death in 1938.

Gradually business improved. The first item to be manufactured in quantity with a slide fastener was a money belt developed in 1917 by a tailor in New York; it was popular among sailors in World War I. Most of the 24,000 fasteners the company sold that year were for such belts. In 1918 the fasteners were incorporated in some 10,000 flying suits made for the Navy.

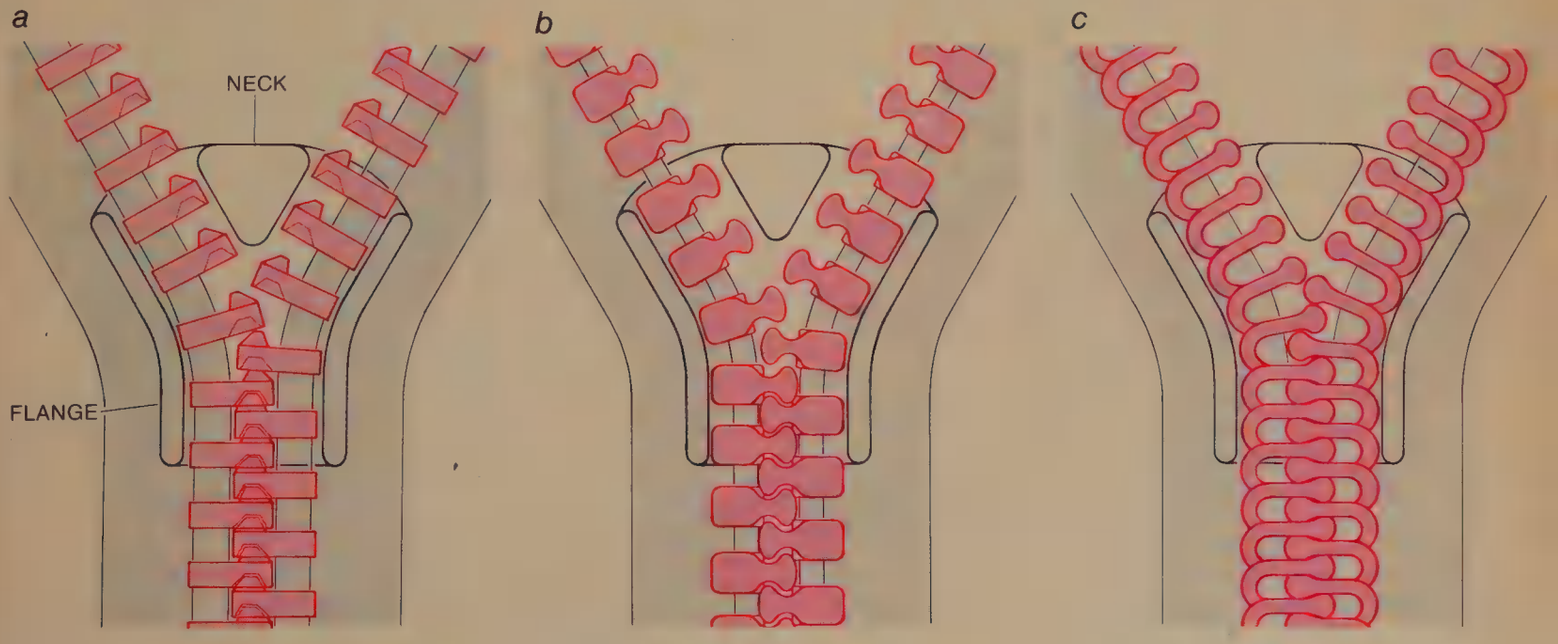
Soon after the war the introduction of slide fasteners on gloves and tobacco pouches helped to make the device familiar to civilian consumers. Probably the biggest impetus, however, came from the introduction of slide fasteners on galoshes in 1923 by the B. F. Goodrich Company, which also originated the name zipper and registered it as a trademark for the Goodrich line of rubber footwear.

Expansion was now rapid. Hookless Fastener, which was the sole maker of slide fasteners in the U.S. from 1917 to 1926, saw its sales rise from 24,000 in 1917 to more than 60 million in 1934. By that time the company had changed the name of its fasteners to Talon, and in 1937 it also changed its corporate name to Talon, Inc. The U.S. production of zippers peaked at about 2.3 billion some 10 years ago; it has declined since then as a result of competition from abroad and stood at less than 1.8 billion in 1981.

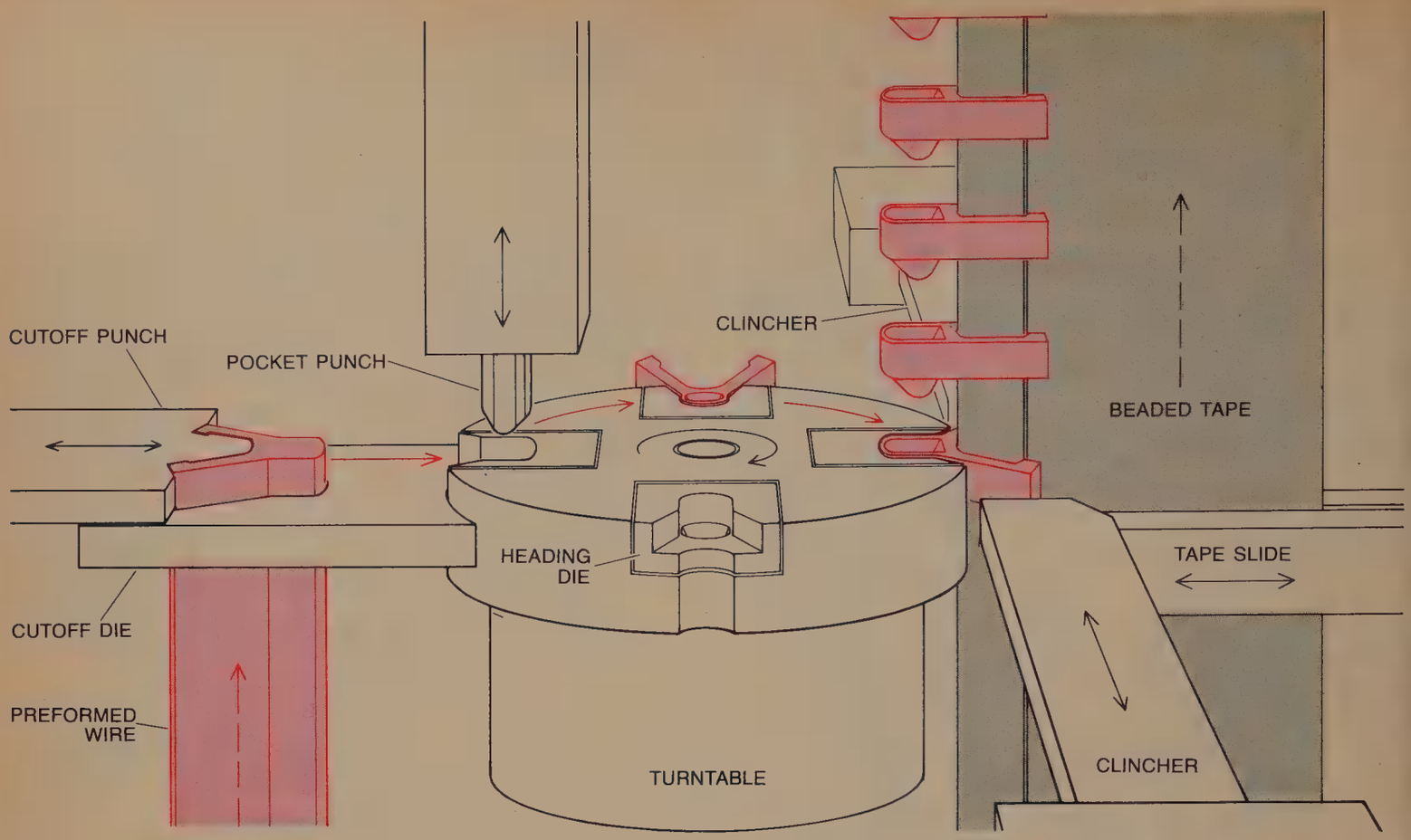
The basic components of a metallic zipper are the teeth, the tape, the "hardware" and the slider. The teeth are the individual units, arrayed on a piece of tape. The tape fitted with teeth is called a stringer. Two stringers joined side by side make up a chain. The hardware consists of parts that prevent the slider



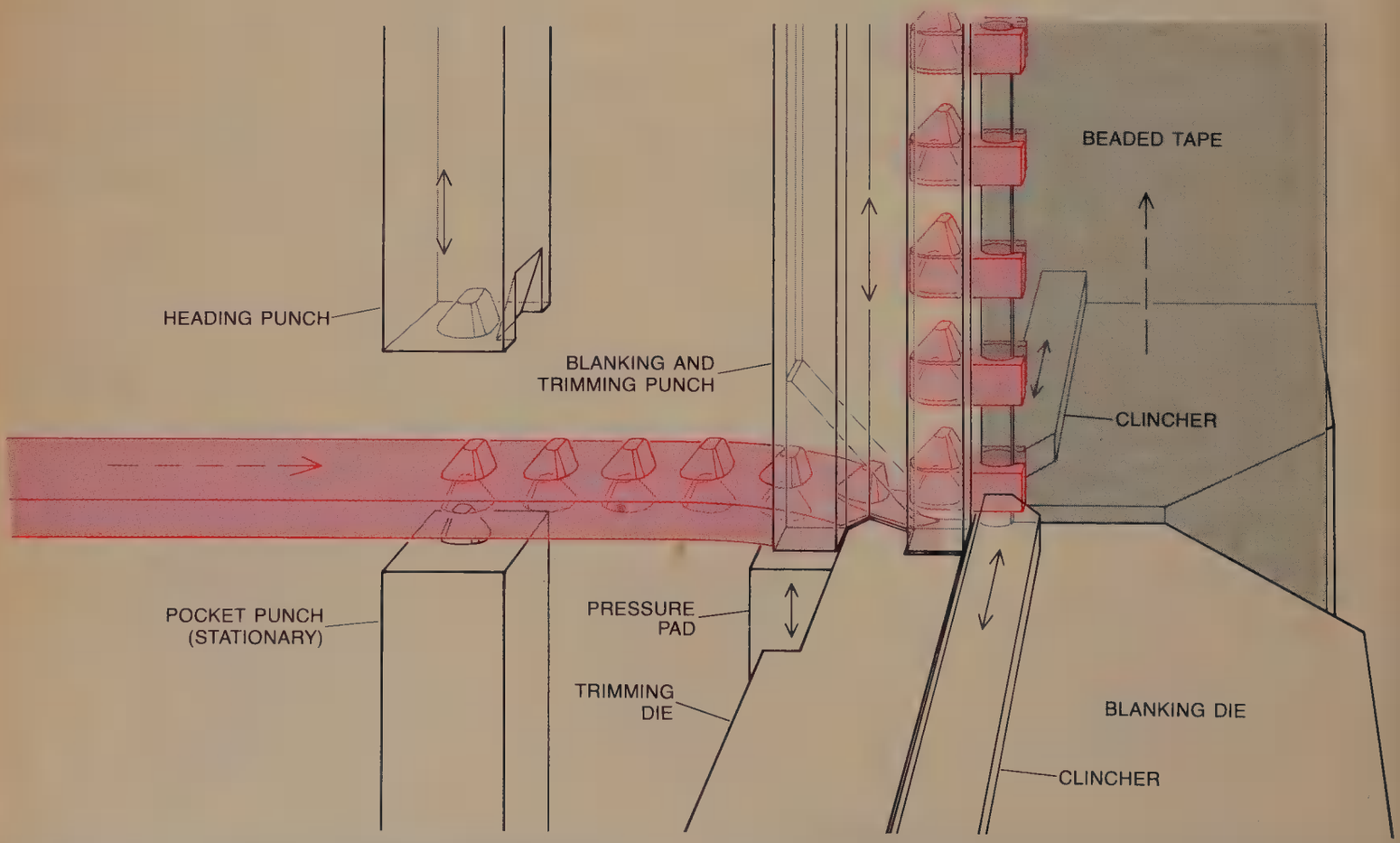
COMPONENTS OF A ZIPPER are portrayed for an open type of the kind put on trousers and for a separating type such as is needed on a jacket. The teeth are sometimes called scoops because a metal tooth in profile resembles a scoop. The fabric tape serves not only to hold the teeth in position but also as the means of sewing the zipper to a garment, a step further served by the tape extension. A tape fitted with teeth or spirals is termed a stringer. Two stringers joined are a chain. A chain provided with a slider and stops is a zipper.



OPERATION OF ZIPPER is shown with the front of the slider removed. In a zipper with metal teeth (*a*) each tooth has a dome or ball on the top of one end and a pocket on the bottom. As the slider is raised its neck positions the teeth at the appropriate angle and distance for the ball of one tooth to fit into the pocket of the opposite tooth; the flanges push the teeth together. In opening a zipper the neck pushes the teeth apart. Plastic spirals or teeth (*b, c*) are made in many different shapes, but the opening and closing principle is the same.



METAL STRINGER is made in several ways, one of which is depicted here. The process begins with wire that has been preformed into a *Y* profile (*left*). As the wire is moved upward in the machine a cutoff punch slices off what will become individual teeth. A sliced piece is pushed into a heading die that is rotating clockwise. There a pocket punch forms the ball and pocket. After two 90-degree turns of the die the piece reaches the tape, which is pushed between the legs of the *Y*. The legs are clamped on the beaded tape by the clincher and then the tape advances. The bead, which does not appear on all tapes, provides more material for the legs of the tooth to grip.



ALTERNATIVE PROCESS for making metal zippers operates on a strip of flat wire. The strip is fed into the machine by rollers. A heading punch works against a stationary pocket punch to form the ball and pocket. Then a blanking punch cuts off notches to form the basic *Y* and at the same time cuts off the front tooth. Even before the front scoop is fully off the strip the clinchers bend the legs of the scoop around the cord of the tape. Then the stringer advances into the next position. The machine can make 50 scoops per second.

from slipping out. The chain is closed and opened by the slider.

If you look closely at a zipper tooth made of metal, you will see a tiny dome on the top of the outer end and a corresponding pocket on the bottom. Together they give the unit something of the configuration of a scoop, which is what a zipper tooth is often called. The dome or ball of a scoop fits into the pocket of the corresponding scoop on the opposite side of the chain; this geometry along the length of the chain is what holds the two stringers together when the zipper is closed.

If you now look closely at a slider, you will see that its left and right sides form a flange. If you are able to peer inside the slider at the top, you will see a solid shape in the form of a *V;* it is called the neck. As you close a zipper the inner *V* and the flanges acting together move the scoops of opposing stringers far enough apart so that they are separated and held at an angle allowing the end of each scoop to slip in between two scoops on the opposite side. The flanges of the slider as it moves upward then push the two stringers together, scoop by scoop. When you open a zipper, the flanges hold the scoops at the angle that allows opposing scoops to slip apart, and the descending *V* separates them.

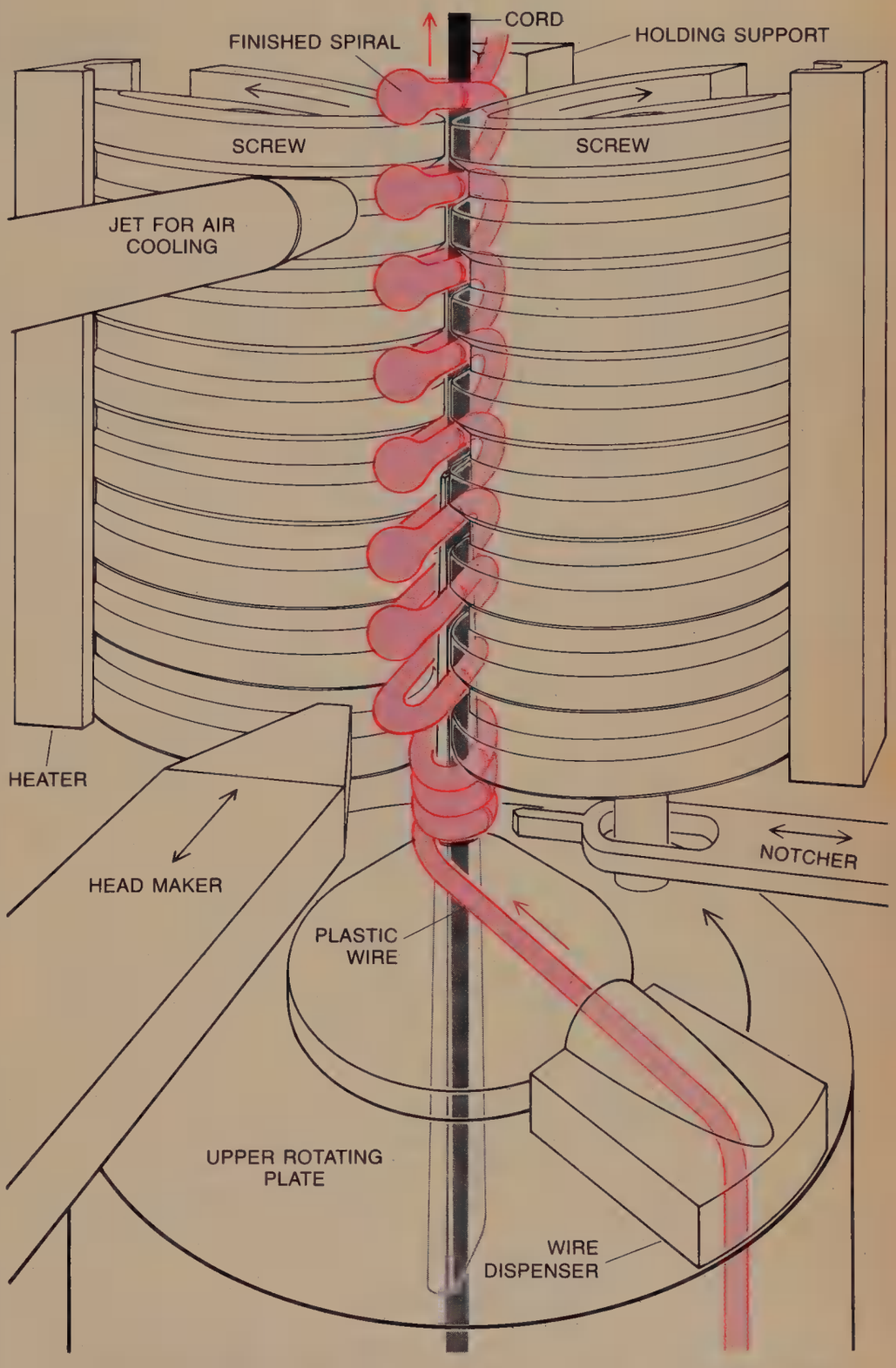
At one time zippers were made only of metal. After World War II the improvement of polymers and of methods for forming them gave rise to a variety of plastic zippers: spirals, coils, ladders and other configurations. In some of them the opposing spirals are held together by geometries that differ from the ball-and-pocket one, but the principle whereby the slider opens and closes the two sides of the chain is much the same.

Slide fasteners are made in various ways because certain types serve better than others in a specific application. For example, the zipper on a dress must be pliable and not likely to irritate the skin, and a plastic zipper in which the stringers consist of fine spirals seems to be the best choice. On the other hand, a zipper on heavy luggage must be strong in order to withstand substantial loads. A large zipper with metallic scoops or plastic coils or teeth is the best.

The size of a chain is identified by a measurement across the joined scoops or spirals. The numbers for metallic chains range from 2 (.135 inch) to 7 (.35 inch). For plastic chains the numbers range from 3 to 9, which correspond roughly to millimeters.

A zipper chain is made in one continuous length and is later cut up into individual zippers, to each of which are added a slider and appropriate stops to limit the travel of the slider according to the type of zipper. A separating zipper, as on a warm-up jacket, has a stop at the top of each stringer and a box and pin at the bottom. An open-top zipper, as on a pair of trousers, has a single stop at the bottom of the chain and a stop at the top of each stringer.

Let us consider the manufacture of zippers in two categories, according to whether the fastening element is made of metal or of plastic. I shall describe three methods of making a metallic chain. In each case a fully automatic machine does the job, and one operator can tend several machines. Essentially what happens in the first two methods is that individual scoops are made in rapid



COILER USING ONE ROUND WIRE makes a modified spiral for a plastic stringer from polyester or nylon wire. Two joined spirals are sewn simultaneously on two tapes. Here one sees the upper part of the process, beginning where the plastic wire emerges from the dispenser and is wound around a tapered mandrel by the action of a rotating plate. At the same time a cord is pulled up through the center of the spiral; it helps to secure the spiral to the tape. As the spiral advances up the mandrel it is notched from the side; the notches also help to secure the spiral to the tape. Two heated screws, rotating in opposite directions, seize the plastic and stretch it so that the front part becomes horizontal and properly spaced. The front of the spiral is compressed to make the head. The spiral is heated for a period of time at a set temperature and is then quickly cooled so that it will maintain the desired shape. This is the process called thermosetting. Left-hand and right-hand units work side by side to make matching stringers.

succession and clamped onto an unrolling ribbon of fabric tape. The edge of the tape is beaded so that the legs of the scoop grip firmly.

The first process begins with a round wire made of brass, aluminum or a nickel-silver alloy. The wire goes through several passes in a rolling mill that form it into a *Y* profile, so that as it is later sliced like a loaf of bread a series of individual *Y*'s is produced. Each of them becomes a scoop; the branches of the *Y* are the legs that clamp to the beaded tape.

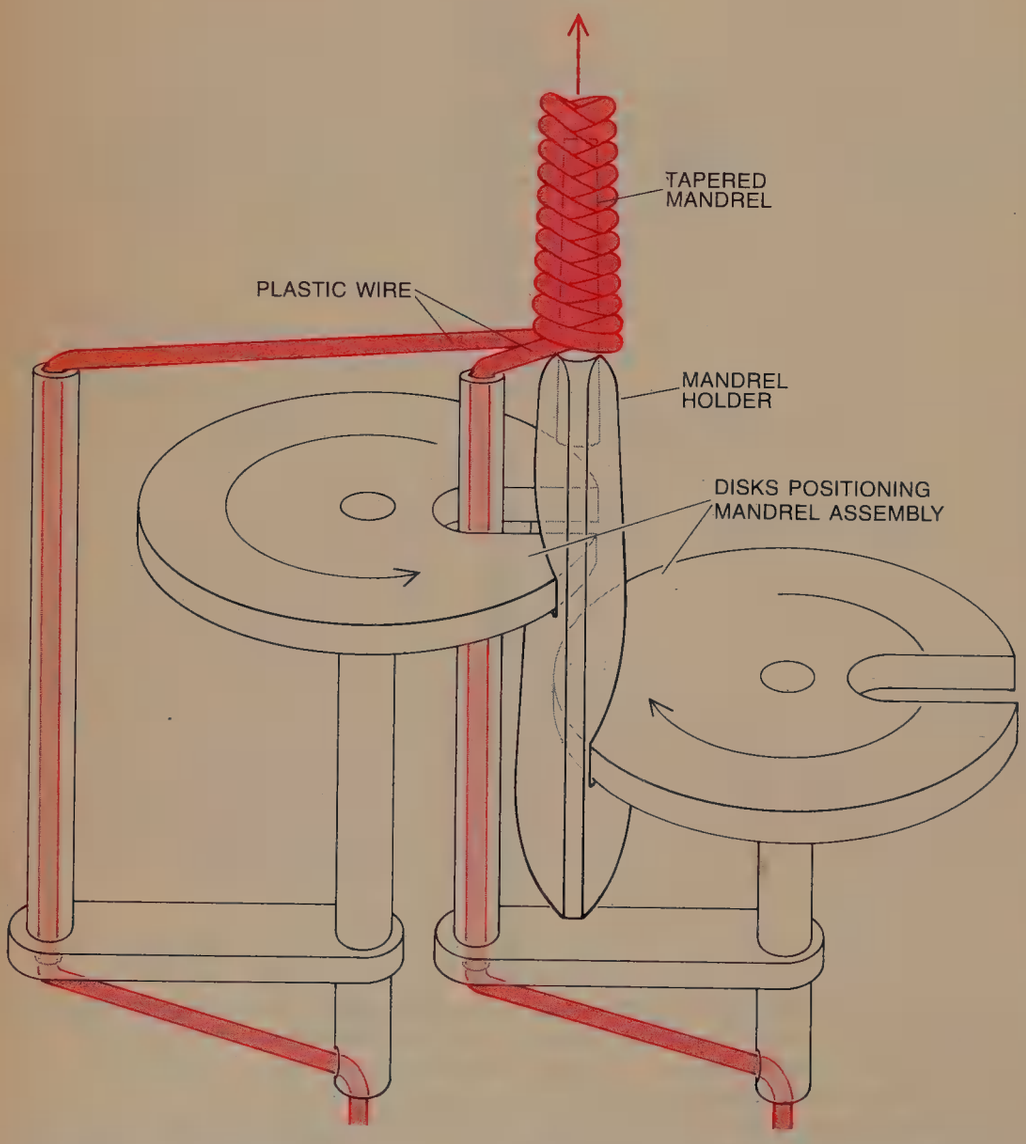
As the profiled wire is fed into the zipper machine a cutting punch slices a blank *Y* and pushes it into a heading die, one of four dies on a turntable. A pocket punch descends on the *Y* and forms the ball and pocket. After two 90-degree rotations of the turntable, during each of which a blank is added to another die, the blank reaches the beaded tape. The beaded edge of the tape is pushed into the scoop, and clinchers close the legs of the scoop around the bead. The tape is then raised by approximately twice the thickness of the scoop plus about 10 percent, a distance calculated to make room for the next scoop and also for the companion scoop from the other side that will slip between them when the zipper is closed.

This method of making a metal zipper chain is some 60 years old. Indeed, the machine is a direct descendant of the first automatic chain-making machine, which was invented by Sundback in 1923. Before that time the general method of making zippers was to stamp out scoops individually. The scoops were tumbled to eliminate sharp edges, plated and inserted manually into a fixture. When the fixture was filled, a corded tape was threaded through the legs of the scoops. The legs were then closed on the tape by a power press. Later the manual procedure was mechanized, but the process was still slow and unreliable. Moreover, it was wasteful: the stamping of the scoops left some 40 percent scrap. Sundback's automatic chain-making machine left no scrap, and it was a tremendous success. Round-wire processes, however, are relatively slow and therefore not much used today.

In the second process, which is traced back to a patent received in 1940 by Frederick Ulrich of the Conmar Products Corporation, the machine works not on a round wire but on a flattened strip of wire. Two rollers feed the wire into the machine, where a heading punch works against a stationary pocket punch to form the head and pocket on what will become the individual tooth of the stringer. Next a blanking punch cuts off notches on both sides of the wire and at the same time cuts the front scoop from the flattened wire. Before the blank is fully cut its legs are bent around the bead of the tape. A well-designed machine of this kind operates at a rate of about 50 scoops per second.

In the third process, which originated with a patent received in 1932 by Gustav Johnson, the teeth are not stamped but are cast directly on the tape. The tape moves through an open two-part mold, which has a series of cavities corresponding to the teeth. When the mold is closed, molten zinc is injected into it under pressure. The mold is water-cooled, and when the molten metal solidifies, the mold opens. The tape with the teeth moves into the next position and the dross is cut off.

Once a metallic stringer has been made by any of these three processes it is put through certain finishing operations. Two stringers are joined into a continuous chain by a closing fixture that resembles the slider on a standard zipper. The closed chain passes through rollers that press it to the proper thickness, and then it goes through a series of wire brushes that remove sharp edges. The next station is a tank of starch. Emerging from that, the chain is wrung out by squeezing rollers and passed over hot drying cans. The purpose of this operation is to make the tape flat. Finally the scoops receive an application of hot wax, which makes the slider operate smoothly when the zipper is new. (When the zipper has been broken in, the slider will operate smoothly without wax.) The continuous finished chain is rolled onto a spool and is ready to be assembled into a zipper.



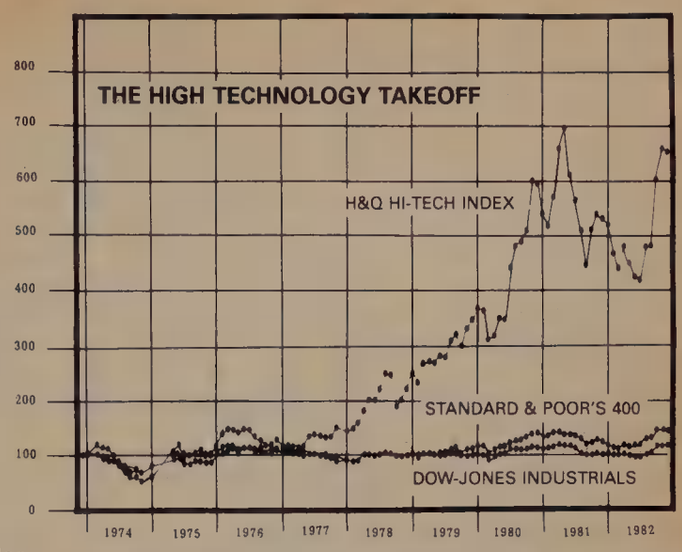
COILER FOR TWO WIRES winds them on one mandrel. This double-wire coiler operates on plastic wire that is profiled rather than round. The wires are wound in opposite directions on a tapered mandrel. After the plastic coil is removed from the mandrel and thermoset on another wheel it is woven, each coil separately, into the stringer. The method is sparing of material.

The plastic zipper of today made its first appearance in Germany after World War II. The German zipper manufacturers had lost factories and equipment during the war, and what had survived was antiquated. Faced with the prospect of having to rebuild their industry, the Germans decided to develop plastic zippers.

The pioneering company was the West German firm Opti-Werk GMBH. As always, however, the development of such zippers was the work of many inventors. Among them were two Ameri-

Now Available . . . a comprehensive journal for investors or anyone interested in high technology

Investor's Guide To High Technology Corporations



What is the Guide?

The Investor's Guide is a new, quarterly journal of information, corporate profiles, investment newsletters, and general overviews of high technology developments. Each issue reports on a variety of *established, emerging,* and *embryonic* public corporations.

All material is presented in a sturdy, three-ring binder, making it convenient to add or subtract information according to each subscriber's needs. The June 1983 issue contains 150 pages. There are two editions—one for private investors, one for corporations.

Major Industries Surveyed

- Aerospace
- Biotechnology/ Genetic Engineering
- Chemicals
- Computer Hardware
- Computer Software and Services
- Computer Peripherals
- Electronics (General)
- Environmental Control
- Fiber Optics
- Laser Technology
- Materials Science
- Medical Devices
- Pharmaceuticals
- Robotics
- Semiconductors
- Telecommunications

Hi-Tech Investment Newsletters

Each issue of the Guide presents three separate newsletters with suggested stocks for different investor types:

CONSERVATIVE ADVISOR

Directed to cautious investors who look for established corporations in the high technology field with good track records and, in most cases, moderate-to-high quarterly dividends. Risk to investors is low. Long-term purchases are generally recommended.

MODERATE ADVISOR

Directed to investors willing to take moderate risk in emerging corporations, usually from 2-9 years old. Long-term purchases are generally recommended for investors willing to grow with these emerging companies over the years, despite the possibility of temporary setbacks.

SPECULATIVE ADVISOR

Directed to investors willing to assume higher risk in new corporations with little or no records, but with potential to bloom (quickly or otherwise) into substantial corporations.

(Investors should be aware that, even though the stock market has trended upward since 1897, there are periods of economic recession that depress the value of many stock portfolios. New investors are cautioned to read the introduction in the Guide very carefully.)

Corporate Profiles

Each corporation capsulized in the investment newsletters is profiled in greater depth, with some 20 categories of information, including:

- Major activities in product or service commercialization
- R & D commitments
- Stock price history
- Revenues and assets
- Liabilities
- Management statements

Hi-Tech Products & Services

The first issue contains a comprehensive listing of products and services produced by the corporations monitored by AIIS. Subsequent issues will contain glossaries and cross-references.

News Briefs and Industry Trends

The Guide provides concise reports of the high technology field, investment news, and projected growth trends for selected industries.

Corporate Edition

In addition to the material presented in the private investor edition, the corporate edition contains information of interest to money managers in small, medium, and large corporations that hold or are planning to hold stock portfolios.

Since favorable dividend tax credits exist for corporations owning preferred stock (currently offering a greater return than bonds and money market funds), the first issue highlights some of those stocks paying attractive dividends.

Supplements

For annual subscribers, AIIS will publish occasional supplements containing updated reports, information about new offerings, and other topical material.

The growth of high technology corporations is evidenced by the chart prepared by the investment banking firm of Hambrecht & Quist, whose index of some 100 hi-tech stocks greatly outperforms the major market indexes. For the purpose of comparison, all indexes have been set at 100 as of December 1973. The H&Q Hi-Tech index at the end of March 1983 stood at 770.74, well over 100 points higher than at the close of 1982.

New Hi-Tech Companies

Each issue provides a comprehensive listing of new issues in high technology filed with the Securities and Exchange Commission.

For New Investors

For beginning or recent investors, the Guide provides a clearly written introduction to stock market investing, which includes an explanation of the stock market exchanges, discount and regular brokerage services, terms and abbreviations, stock price-watch charts, and other useful information.

AIIS

The Guide is published by American Investor Information Services Inc., a Pennsylvania corporation registered with the U.S. Securities & Exchange Commission under the Investment Adviser's Act of 1940. AIIS staff members and independent consultants contribute information to the Publisher who is solely responsible for investment suggestions.

Sources utilized include SEC reports, technology and financial journals, corporation reports, communication with management and specialists, and Standard & Poor's reports. Additional data are received through computer link-up with the Dow Jones News Retrieval Service.

American Investor Information Services, Inc.
1627 Spruce Street, Philadelphia, PA 19103

SA-J

Subscription Fee Schedules

For private investors
The annual rate is $150.00. A sample issue (June 1983, 150 pages) is available for $59.95. Binder included. Supplements not included in sample issue.

Tax-deductible when used as an aid in the management of your investments.

For corporations
The annual rate for the corporate edition is $295.00. (Additional subscriptions to same address at 50% annual rate.) A sample issue (June 1983) is available for $100.00.

For 30-day billing, please send order on company stationery or purchase order.

Please enter my subscription to the Investor's Guide.

☐ Payment enclosed for one year ($150.00) or for sample issue ($59.95). Make checks payable to American Investor Inc. Pa. residents add 6% sales tax.

☐ Charge my VISA or MASTERCARD (circle one):

Card # _____ Exp. Date _____

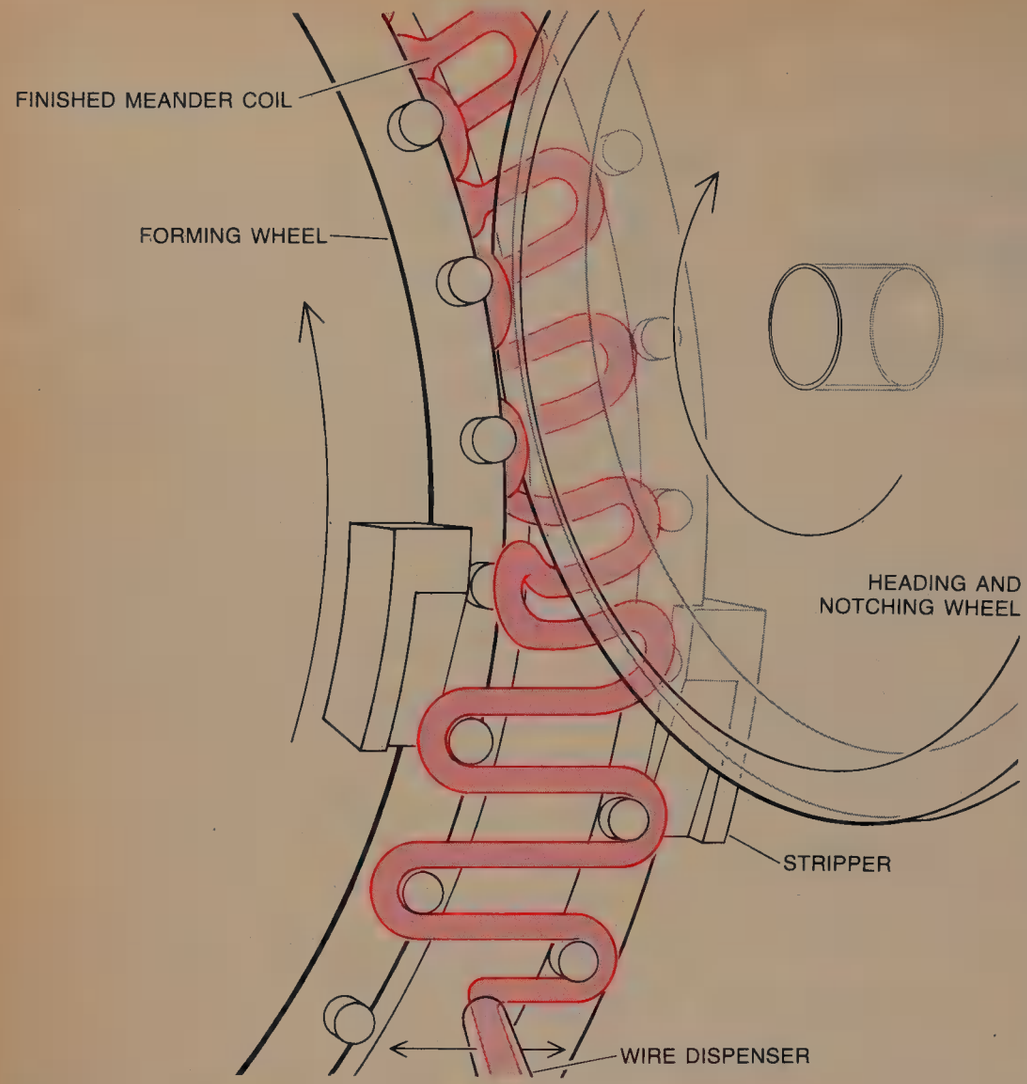
NAME _____

ADDRESS _____

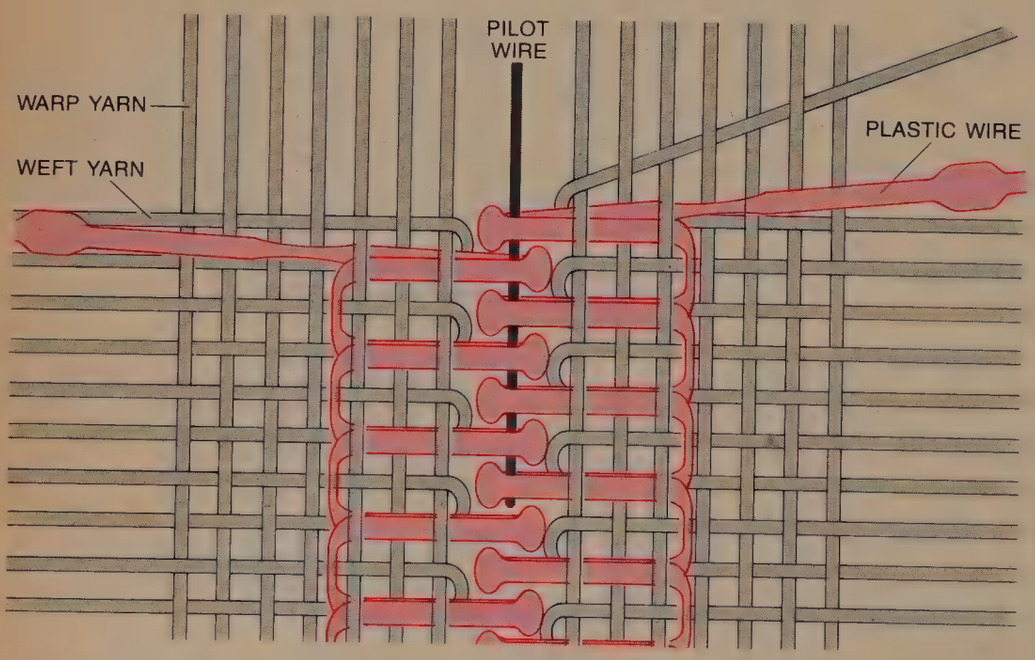
CITY, STATE, ZIP _____

For more information, write to AIIS or call (215) 732-5350

VISA/MASTERCARD holders, orders only: CALL TOLL FREE—24 Hours, 7 days a week
800-257-5858 (In New Jersey, call 609-235-4747)



ROUND PLASTIC WIRE is formed into a ladder (or meander) coil by a pair of wheels. The wire dispenser moves intermittently and crosswise to the forming wheel, and the forming wheel also moves intermittently so that the wire is laid between the pins. Stationary strippers lift the wire from the pins. The second wheel presses the ladder formation into a *U* on the forming wheel, the plastic is thermoset and the heads are formed. Finally the coils are sewn on the tapes.



LOOM WEAVING of notched plastic wire into a zipper chain is depicted. Two shuttles hold yarn and two hold notched round plastic wire. The pilot wire is attached to the frame of the loom, and as the weaving progresses it slips out of the chain. Chain made in this way is of the best quality and is particularly suitable for garments. The operation is slow, because only one of the four shuttles works while the other three are waiting, and so the manufacturing cost is quite high. Loom weaving is not done in the U.S., but many zippers of this type are imported.

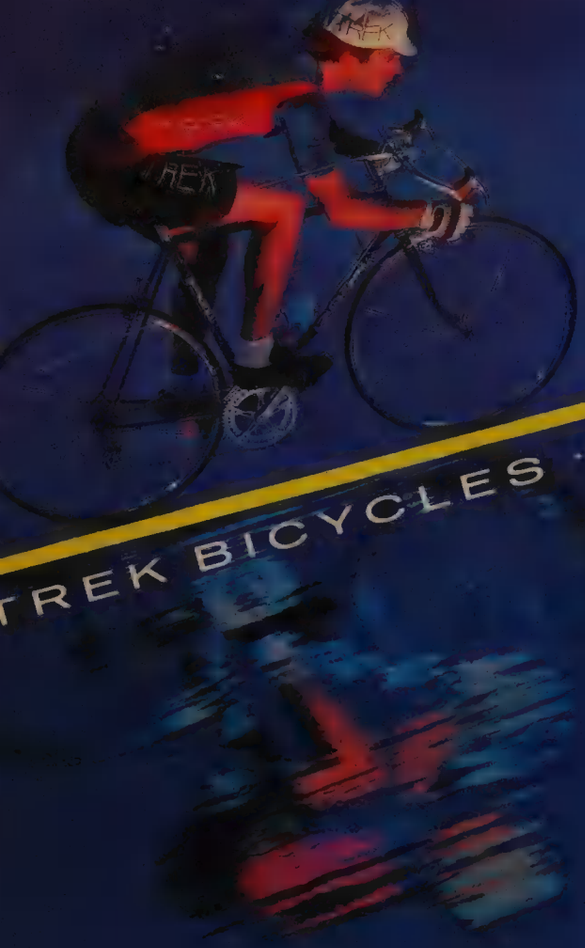
cans. In 1942 Alden W. Hanson received a patent on a method of sewing a plastic coil on a tape; in 1951 Nicholas A. Wahl received one on a method of winding two profiled plastic wires on a single mandrel. Thereafter J. R. Ruhrman and his associates were awarded a German patent on a plastic ladder chain. A. Gerlach and his associates, and independently the firm William Prym–Werke, were awarded patents on a notched plastic wire that could be woven into a tape. Another important patent was granted in 1968 to the Australian inventor E. E. Cuckson and his associates for a system that would continuously cast scoops in plastic zippers. A coiler that folds two round plastic wires into a thermosetting wheel was invented in the U.S. in 1969.

A plastic stringer has spirals made of polyester or nylon formed as a modified spiral, a meander or other shapes. In most zippers today the plastic spirals are sewn or woven on the tape, except in one version where the teeth are cast directly on the tape. Both polyester and nylon have a melting point above 204 degrees Celsius (400 degrees Fahrenheit), so that a garment with a plastic zipper can be safely ironed.

For the purpose of this discussion I shall describe two of the machines that make plastic zipper chain. One is the machine that winds round plastic wire on a mandrel. Its essential parts are the stationary tapered mandrel and a stationary cord section, both mounted on an offset shaft. (The cord runs vertically through the spiral of the stringer; it helps to secure the spiral to the tape.) A rotating spool pays out the plastic wire, and the cord is pulled from a cone. Other important parts are two heated screws, a notcher and a head maker.

The plastic wire comes off the rotating spool and passes through two rotating plates and into a dispenser, which winds it on the mandrel. At the same time the cord is pulled through the offset shaft and a slot in the mandrel, so that it ends up passing through the center of the spiral. The cord is not always used. With each revolution of the dispenser the spiral advances up the mandrel. As it advances it is notched from the side; the notches are another means of helping to secure the spirals when they are sewn on the tape.

The two heated screws, rotating in opposite directions toward the notched spiral, seize it and stretch it so that its front end is horizontal and properly spaced. A stationary bar behind the screws holds the spiral in the desired position. The screws are heated so that they set the thermosetting plastic of the spiral. After the stretching the front of the spiral is compressed to make the heads. At the end of its travel the spiral is cooled by a jet of air. Often the screws are replaced by gears to stretch



TREK BICYCLES · BRILLIANTLY RESPONSIVE.

The first criterion in defining bicycle performance is frame design. At Trek, we take bicycle frames seriously.

Designed by craftsmen who forged their reputations building custom framesets, every Trek bicycle frame is tailored to a specific riding need. This Function-Specific™ design redefines the relationship of rider to machine, providing the crisp, exhilarating experience of total response.

Trek frames are crafted of high strength-to-weight chrome or manganese molybdenum double-butted tubing.

Trek then matches components from around the world to complement the Function-Specific™ design of these outstanding frames.

The result – uncommon bicycles of exceptional strength and resiliency. A synthesis of meticulous design, esoteric materials and American craftsmanship.

Trek bicycles transcend the commonplace.

For the name of your nearest Trek dealer call Toll Free 800-558-TREK Ext. 400 (In Wisconsin call 414-478-2191)

TREK®

AMERICAN HANDBUILT BICYCLES

TREK BICYCLE CORPORATION
WATERLOO, WISCONSIN 53594

Authorized Trek Dealers

This is a partial listing of Trek dealers. If there is no dealer in your area, call 800-558-TREK. (In Wisconsin call 414-478-2195.)

abama
burn/Freewheeler 887-8117
rmingham/Cahaba Cycles 969-0476

izona
oenix/Happiness House Schwinn Cyclery 995-8474
oenix/Landis Cyclery 264-5681
mpe/Landis Cyclery 839-9383
cson/Bicycles West 887-7770

lifornia
aheim/Anaheim Schwinn Cyclery 535-3231
rkeley/Missing Link 843-7471
nita/Bonita Schwinn Cyclery 475-7205
marillo/Open Air Bicycles 484-4433
remont/Bud's Bike Shop 626-3285
llerton/Fullerton Bike Shop 879-8310
leta/Open Air Bicycles 967-9669
ntington Beach/Two Wheel Transit Authority 536-6940
a Vista/Open Air Bicycles 968-5571
guna Beach/Laguna Beach Cyclery 494-1522
s Altos/Bicycle Outfitter 948-8092
s Angeles/I. Martin Imports Bicycles 653-6900
rthridge/Cycle World 349-6900
lo Alto/Palo Alto Bikes 328-7411
rt Hueneme/Open Air Bicycles 985-5045
dwood City/Chain Reaction Bicycles 366-7130
cramento/City Sport Works 332-6453
n Francisco/Presidio Bicycle Shop 751-3200
nta Barbara/Open Air Bicycles 963-3717
nta Rosa/The Cycle Path Ltd. 526-7121
ntura/Open Air Bicycles 653-1100

lorado
spen/Hub of Aspen 925-7970
ulder/High Wheeler 442-5588
lorado Springs/Criterium 475-0149
nver/Denver Spoke 777-1720
nver/Life Cycle 572-8405
nver/Turin Bicycles, Ltd. 837-1857
rango/Four Faces 247-0500
tes Park/Cosmic Wheels Cyclery 586-2975
rt Collins/The Bike Broker 484-2453

nnecticut
omfield/Bloomfield Bike & Repair 242-9884
d Greenwich/Buzz's Bicycle Shop 637-1665

Delaware
Newark/Two Wheeled Cycle 368-2685

Florida
Orlando/Orange Cycle Works 422-5552
South Miami/Bicycle Store of South Miami 665-0125

Georgia
Athens/Dixon's Bicycling Center, Inc. 549-2453
Atlanta/Vinings Schwinn 952-7731
Dunwoody/Dunwoody Schwinn 455-3171

Idaho
Boise/George's Lightweight Cycles 343-3782

Illinois
Arlington Heights/Bikes Plus 398-1650
Chicago/Village Cycle 751-2412
Evanston/Turin Bicycle Co-op 864-7660
Niles/Spoke 'N Pedal 692-4240
Oak Park/Oak Park Cyclery 386-0282

Indiana
Bloomington/Bicycle Garage 339-3457
Fort Wayne/Denny's Bike Shop 432-7636
West Lafayette/Hodson's Bay Company 743-3506

Iowa
Davenport/World of Bikes 323-8227
Des Moines/Bike World 255-7047
Iowa City/International World of Bikes 351-8337

Kansas
Manhattan/Pathfinder 539-5639
Overland Park/The Bike Rack 642-6115
Wichita/Bicycle Peddler 262-8029

Louisiana
New Orleans/Bikesmith, Inc. 891-2453
New Orleans/Gus Betat & Son 821-6386
Shreveport/Sunset Bicycle 635-5079

Maryland
Gaithersburg/Adam's Bicycles of Gaithersburg 840-0484

Massachusetts
Fall River/Paul's Schwinn 674-6343
Fitchburg/Gamache's Cyclery 343-3140
Worcester/O'Neil's Bicycle Shop 798-0084

Michigan
Ann Arbor/Great Lakes Cycling Center 663-1604
Kalamazoo/Alfred E. Bike 349-9423

Minnesota
Bloomington/Penn Cycle 888-1427
Hopkins/Now Sports 827-2838
Lake Elmo/North Country, Inc. 739-3500
Mankato/The Bike Shop 625-2453
Minneapolis/Freewheel Bike Co-op 339-2219
Richfield/Penn Cycle 866-7540
St. Paul/Now Sports 644-2354

Missouri
Bridgeton/Touring Cyclist 739-5183
Florissant/Touring Cyclist 921-1717
Maplewood/Maplewood Bicycle Sales & Service 781-9566
Richmond Heights/Touring Cyclist 781-7973

Montana
Billings/The Spoke Shop 656-8342

Nebraska
Lincoln/Cycle Tech, Inc. 466-2921
Omaha/Bicycle Specialties 556-2453

New Hampshire
Durham/Durham Bike 868-5634

New Jersey
Brielle/The Brielle Cyclery 528-9121
Montclair/Montclair Bikery 744-7252
Morristown/Marty's Reliable Cycle Center 538-7773
Ridgewood/Ridgewood Cycle Shop 444-2553

New Mexico
Albuquerque/Albuquerque Bike Co-op 265-5170
Los Alamos/Strings 'N Spokes 662-3456

New York
Brooklyn/Sheepshead Bicycle Shop 646-9430
Mt. Kisco/Bicycle World 666-4044
New York/Bicycle Habitat 691-2783
Rochester/Towner's 271-4453

North Carolina
Asheville/Liberty Corner Bicycles 253-7092

Ohio
Centerville/International Pro Bike Shop 433-6687
Columbus/Cross Country Cycle 294-4343
Findlay/The Bicycle Shop 423-2729
Yellow Springs/Yellow Springs Bike Shop 767-1956

Oklahoma
Oklahoma City/Pro Bike, Inc. 721-6707

Oregon
Eugene/Collins Cycle Shop 342-4878
Portland/The Bike Gallery 281-9800

Pennsylvania
Horsham/Pedal Pusher Ski & Sport 672-0202
State College/The Ski Station 237-2655

Texas
Austin/World Cycle 477-2142
Dallas/University Cycle 368-2035
Fort Worth/Bikes Etc. 927-2421
Lewisville/Lewisville Bicycle Center 221-2744

Utah
Salt Lake City/Joe Fisher's Bike Shop 466-3971
Salt Lake City/Sid's Ski & Cycle 485-3204

Virginia
Norfolk/Colley Avenue Bike Shop 622-0006

Washington
Seattle/Angle Lake Cyclery 878-7457
Seattle/Aurora Cycle 783-1000
Seattle/Northwest Cycle 329-2453
Spokane/Wheel Sport 326-3977

Wisconsin
Appleton/Janus 731-2253
Green Bay/Wheels & Boards 468-4745
Hales Corners/Wheel & Sprocket 425-7930
Madison/Yellow Jersey 257-4737
Milwaukee/Rainbow Jersey Bicycle Shop 961-1198
Monona/Village Pedaler 221-0311
Sun Prairie/Bicycle World 837-3440

Wyoming
Lander/Freewheel Sports 332-6616

Canada
Toronto/Bloor Cycle 536-9718

TREK®

How to seek les

There are more people in more places doing more things with Apples than with any other personal computer in the world.

Which covers a broader range of activities than most people care to deal with. So we'd like to detail just one of the many powers an Apple® Personal Computer can put in your hands.

It's called "data base management," a computer term for a human activity which, if you've ever looked up a word in a dictionary or a number in a phone book, you're quite familiar with.

Faster ways to get at the facts.

Data base management is the most common application for computers of all types and vital to the world's governments and major corporations. But since the advent of Apples, it's also become a powerful personal tool.

Because when you turn your paper files into electronic files with an Apple you can look at what you need to know in ways you never knew you could.

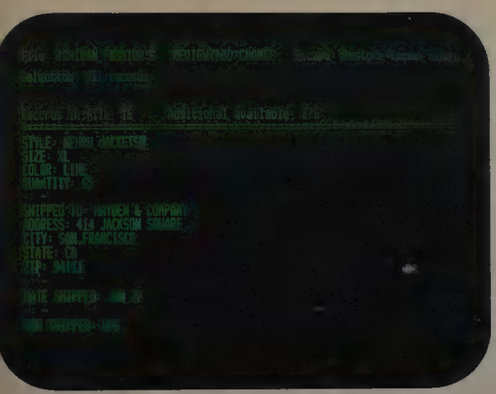
Paper files are limited to a single organizational scheme—either numbering or alphabetizing.

But electronic files can be automatically cross-tabulated,

and find more.

so you can get any information you want, in any order you specify, and get it a lot faster from an Apple than from a file cabinet.

You can selectively retrieve your information—like the name and address of your top salesperson on the East Coast. Or you can generate a



With Apple's Quick File software, you can easily create electronic files that match the paper files you're used to working with. But that's the end of the similarity because...

variety of reports—like a listing of all your accounts ranked alphabetically, or by profitability, size, age, or any other way you specify.

Give yourself instant recall.

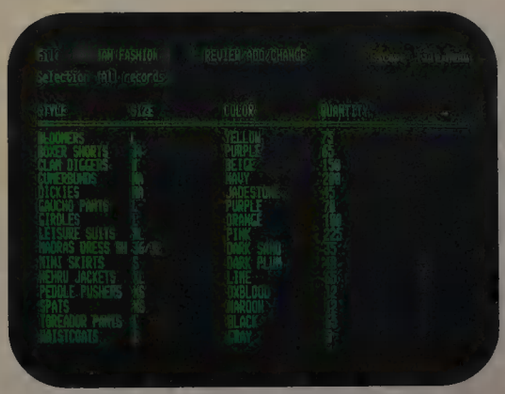
It's easy to understand the value of data base management to the IRS or the Pentagon. But what can it do for you?

Yes, it can shrink your paper files. Apple's ProFile™ storage system can hold up to 3000 pages of double-spaced text.

Yes, it can put the information you need at your fingertips—sales records, personnel files, expense reports, account lists, appointment calendars, price lists and inventories.

Those benefits alone have paid for thousands of Apple systems, and provided smaller businesses and departments with efficiencies that were once available only to the largest organizations. But for you, the manager or proprietor, it means much more.

It means increased personal productivity, because you spend less time searching for information and



Not only can you retrieve information selectively in an instant—whether a single name or all the addresses in a certain zip code—but you can also generate a variety of reports instantly, like this alphabetical inventory or...

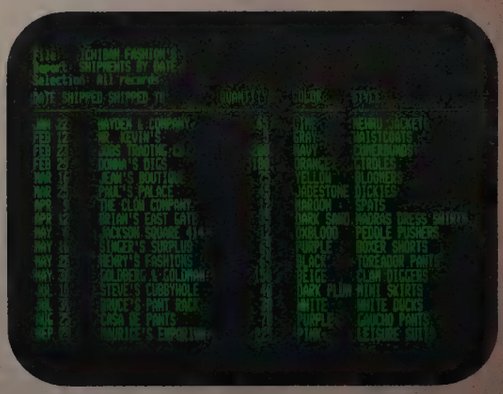
more time deciding how to use it.

It means a totally private information library that you can use as a management tool.

And it means the flexibility to consider information easily in many different ways, allowing you to develop an instinctive feel for the facts and make better decisions.

More ways to establish order.

There are more ready-to-use data management programs available for Apples than for any other personal computer you can buy. Although many are specialized packages for specific types of businesses, Apple



With a few keystrokes, you can look at the same information in chronological order. Or, for that matter, in any other order you want—by location, billing, merchandise sold, or whatever you specify.

offers several data base packages adaptable to any office.

Quick File, for example, is designed for individual use by managers or professionals. It's extremely flexible, easy to use, and well-suited to a wide variety of applications.

Put this in your memory bank.

For your information, Apple has more ways to obtain, store and retrieve information than any other personal computer. We also have more dealers to show you how it works and keep it working—over 1300 of them.

So call the number below to find the authorized Apple dealer nearest you.

And learn how to manipulate the facts to your advantage.

apple®

The most personal computer.

Call (800) 538-9696 for the location of the authorized Apple dealer nearest you, or for information regarding corporate purchases through our National Account Program. In California (800) 662-9238. Or write Apple Computer Inc., Advertising and Promotion Dept., 20525 Mariani Ave., Cupertino, CA 95014. ©1983 Apple Computer Inc.

Our Friends Indeed:

Marlboro Music Salutes Its Corporate And Foundation Supporters And Invites You To Join Them!

RICHARD STOLTZMAN, RONALD LEONARD, ANDRE-MICHEL SCHUB, FELIX GALIMIR

This unique musical community, presided over by Rudolf Serkin, has been described as "perhaps the highest concentration of artistic activity, of whatever kind, that takes place anywhere, anytime, in the United States and possibly the world." The alumni of this internationally acclaimed center for advanced musical studies have taken their places in each succeeding generation of musical leaders where their inspiration and vitality have helped shape our musical lives.

CECILE LICAD, DAVID SOYER, RUDOLF SERKIN

Marlboro reviews with gratitude and pride the list of corporations and foundations which have supported its programs during its first thirty years, and acknowledges the National Endowment for the Arts for its annual support and for the Challenge Grant which played such an important role in the recent Endowment Campaign.

For information about Music from Marlboro concerts or radio broadcasts in your area, Marlboro Recording Society releases or the summer Music Festival in Vermont, please write Marlboro Music, 201 West 54th Street, Suite 4C, New York, New York 10019.

PABLO CASALS

Benjamin and Elizabeth Abrams Foundation, Inc.
The Alford Foundation, Inc.
American Broadcasting Company
Arwood Foundation, Inc.
Ashe Foundation, Inc.
Baer Securities Corporation
The Bell System*
Burlington Industries, Inc.
CBS Foundation
Cargill Incorporated
Cricket Foundation
Charles E. Culpeper Foundation
Dolby Laboratories
The Dover Fund, Inc.
Exxon Corporation
Ford Foundation
Foundation for Advanced Education in the Sciences, Inc.
The Fromm Music Foundation
William Randolph Hearst Foundation
Heineman Foundation for Research, Educational, Charitable and Scientific Purposes
Gunter Henle Verlag
Hochschild Fund
IBM, Essex Junction, Vermont
The Island Foundation
The JDR 3rd Fund, Inc.
L.A.W. Fund, Inc.
The Joe and Emily Lowe Foundation, Inc.
Maya Corporation
The Andrew W. Mellon Foundation
The Charles E. Merrill Trust
Henry and Lucy Moses Fund, Inc.
National Broadcasting Company, Inc.
Walter W. Naumburg Foundation, Inc.
The New York Community Trust
The May and Morris Newburger Foundation
Edward John Noble Foundation
William Penn Foundation
Philip Morris Incorporated
The Presser Foundation
The Rockefeller Foundation
The Martha Baird Rockefeller Foundation for Music, Inc.
Rolled Alloys
The Scherman Foundation, Inc.
Scientific American, Inc.
The Franz W. Sichel Foundation
Spanel Foundation, Inc.
Steinway & Sons
The Florence Louchheim Stol Foundation
The Jean Tennyson Foundation, Inc.
Time, Inc.
von Siemens Fund
The Sulzberger Foundation, Inc.
United Technologies Corporation
Bruno Walter Memorial Foundation
Anonymous Donors

**Special funds given in connection with the Bell Telephone Hour program "Casals at Marlboro."*

and form the spiral and then the thermosetting is done by induction heat.

In a chain of the spiral type the two stringers must have oppositely wound spirals. Therefore two machines are necessary to wind a left coil and a right one. Normally they operate side by side so that the two spirals can be joined into a double spiral. The joined double spiral is ready to be sewn onto the tape.

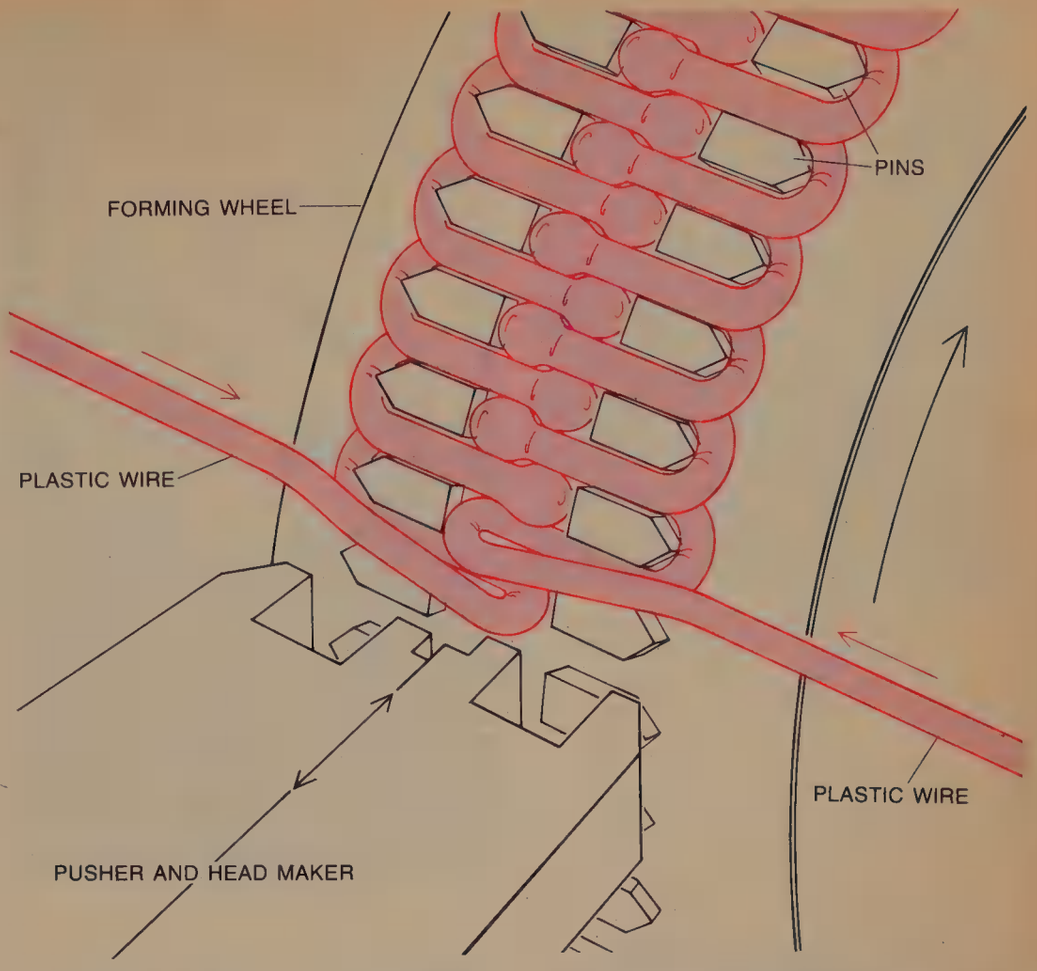
The continuously cast plastic scoops are made on a rotating casting wheel that has along its perimeter uniformly spaced cavities corresponding to the shape of the teeth. A pair of cords are fed to each line of teeth. The cords hold the series of teeth together and facilitate the removal of the string of units from the wheel. Semimolten plastic material is delivered to each passing scoop mold on the rotating wheel. An apron holds the material in place until it solidifies, and then a stationary knife separates the surplus plastic from the units. Next the units go through a folding machine, where they are folded into the proper shape and thermoset. The next step is to sew them on the zipper tape.

A plastic chain goes through finishing operations after it is made. It is put through a starch tank, squeezing rollers and a drying apparatus to make the tape flat. Finally the continuous chain is wound on a paper spool.

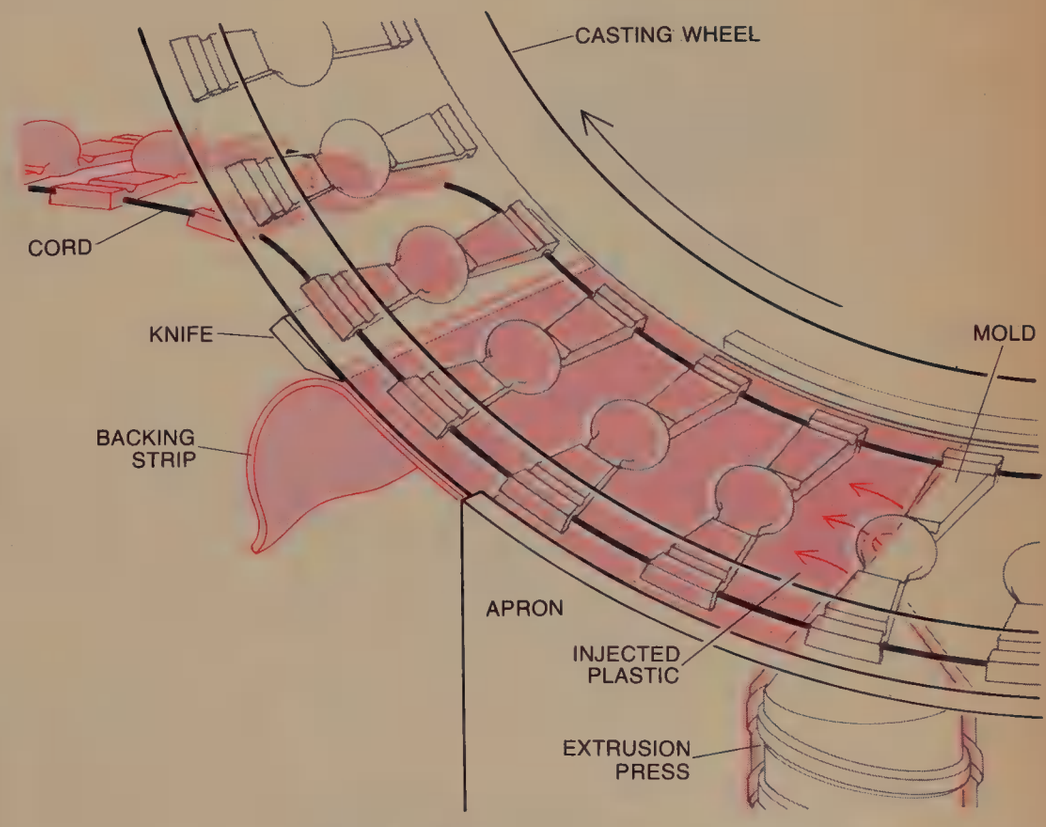
In order to close or open a zipper a slider is needed. The slider body and the pull tab are either stamped from a strip of mild steel or die-cast out of zinc. In general the stamped-out sliders are put on metal chains and the die-cast sliders are put on plastic chains. A slider on a plastic chain must have ribs inside its body to force the spiral into the right position when the zipper is being closed. It is difficult to make the ribs in a stamping process, which is why the die-cast slider is commonest on plastic chains.

To make the slider rustproof the component parts are plated with brass or zinc. An exception appears in what are called automatic sliders. Such a slider has a spring that holds the slider in place until the pull tab is lifted; the spring is likely to be made of stainless steel. Often the slider is enameled to match the color of the chain. The body, the pull tab and (in automatic sliders) the spring are assembled in a special machine.

The last step in making a zipper is to assemble the chain, the slider (two sliders for the type of zipper that can be opened and closed from each end) and the stops. The assembling machinery is different for metallic and plastic zippers, but the methods are essentially the same. A typical assembly operation begins with the continuous zipper chain on a spool. A pair of circular pullers unrolls the chain, pulling it first through a gapper that removes teeth at certain in-



COILER FOLDING TWO ROUND WIRES makes the right and left spirals at the same time. The two round plastic wires are fed from each side into the forming wheel, in a confined space where they form loops. The pusher and head maker force the wire between the pins. After thermosetting the product is taken off the wheel as two closed spirals, which are sewn on the tapes.



CASTING OF PLASTIC TEETH is done on a rotating wheel that has uniformly spaced cavities corresponding to the shape of the teeth. The two cords fed to each line of teeth connect them and make it easier to remove individual units from the wheel. Semimolten plastic is delivered from the extrusion press to the cavities, where it is held in place by the cooled apron until it solidifies. The knife removes the surplus plastic. In a separate operation (not shown) the teeth are folded into the shape of a *U;* they are then ready to be sewn on a flat tape.

Discwasher Wants to Make You a Video Authority

For your copy of "THE VIDEO REFERENCE", complete the coupon below and return it to Discwasher, along with a check or money order for $1 to cover postage and handling.

Please send me _____ copies of "THE VIDEO REFERENCE". Enclosed is my check or money order for $1/copy to cover postage and handling.

Name

Address

City State Zip

discwasher

1407 NORTH PROVIDENCE ROAD, P.O. BOX 6021, DEPT. SA, COLUMBIA, MO 65205 USA

A DIVISION OF JENSEN an ESMARK Company

TELESCOPES

CELESTRON C8 TELESCOPE

Precision Optics From **CELESTRON®**

Celestron offers a complete line of superb quality Telescopes, Spotting Scopes, Binoculars and Telephoto Lenses

- Available in 2.4 to 14-inch diameters
- All price ranges
- Most adapt to 35mm cameras

FREE product, ordering and price information, or send $3.00 for detailed color catalog on selecting/using our products.

CALL TOLL FREE 1-800-421-1526

CELESTRON INTERNATIONAL
P.O. Box 3578-SA, 2835 Columbia St.
Torrance, CA 90503

Name

Address

City State Zip

ANTIQUITY

A Periodical Review of Archaeology
edited by Glyn Daniel

Founded in 1927 by O. G. S. Crawford, ANTIQUITY has appeared regularly ever since and won acclaim the world over as the most authoritative journal in its field. While written by specialists, the articles, notes and reviews are popular in character and indispensable to all interested in the development of man and his past.

Professor Glyn Daniel, Faculty of Archaeology and Anthropology in the University of Cambridge, and Fellow of St John's College, has been Editor since 1956.

The annual subscription, postage included, is $35. Subscription forms and bankers' orders are available on request from

ANTIQUITY PUBLICATIONS LIMITED
Heffers Printers Ltd, King's Hedges Road,
Cambridge, England CB4 2PQ

tervals to provide a stretch of empty tape and simultaneously attaches the bottom stop; when the gap is cut across at the midpoint, it determines the length of the zipper and provides the bit of extra tape needed to properly sew the zipper into a garment. The gap can also be made on the chain machine by omitting the teeth. The chain then goes through a machine that slips the slider through the gap in the chain, and the top stop is put on. The last step is pinking, that is, cutting across the chain in the middle of a gap. Sometimes the operations are combined. The result is an individual slide fastener, ready to be attached to a garment or bag or whatever product it was designed for.

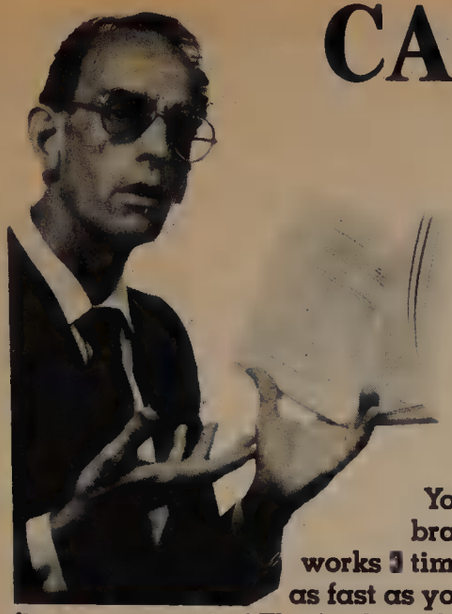
The manufacturing process is slightly different for a separating fastener. A cotton tape, laminated with nylon, is fused across the gap as a reinforcement and top stops are attached. The laminated tape is severed along an extension of the axis of the zipper. The slider is slipped on one stringer and the box is also mounted there. A pin is mounted on the other stringer. When the zipper is opened all the way, the pin pulls out of the box through the slider and the fastener is fully separated, with the slider held on its stringer by the box.

One might suppose little in the way of new developments could be expected in a device as simple and standardized as the zipper. Actually, however, several possibilities are in sight, particularly for plastic zippers. One is for zippers both much larger and much smaller than the ones now on the market. The larger ones would be for heavy-duty action in fastening, say, curtainlike dividers that allow small meeting places to be made in a large room. The smaller zippers would be for products that reflect the current trend toward miniaturization, such as for carrying cases to fit the Walkman type of portable radio.

Another prospect is for an inexpensive zipper that would serve on disposable paper garments and other objects designed to be used once and then discarded. In a more exotic application one can envision a zipper that is strong, airtight and yet reasonably flexible to be put in garments for astronauts. If such a zipper is developed, it would probably turn up soon in consumer products that need fasteners with those attributes.

A final example is an application that is still looking for a zipper. In surgery there is a need for an airtight and chemically inert zipper that would serve in place of stitches in an incision that might have to be reopened, say, for access to an artificial organ such as a heart pacemaker. Several surgeons have expressed a desire for such a device; to design and make it will be a challenge for medical investigators and the zipper industry.

CALCULATOR MATH?



Your brain works 3 times as fast as your fingers can write! That's where the Calculator comes in — you can go three times as fast with your machine doing the dirty work.

It's easier and more interesting. And you'll understand it better when it's fun. (If you haven't yet tried out a modern scientific-statistical calculator, you're in for a pleasant surprise — see below!)

WHY MATH? To get a promotion or a new job. Just to hang on to that job you've got. To quicken your mind. To understand this world you live in. To impress people. Just for some fun you'll feel good about. (Read those letters below to see how it has worked for others.)

Do it at home... You can start right away and go as fast as you like! Put it aside while you're busy with something else — in short, *set your own pace!*

Lots of us like to avoid being compared with the other members of a class — we like to set our own standards. Too, we like to save the time and expense of going to a school or seminar. (Best of all, it's a lot cheaper to study at home where the rent is already paid — *read on!*)

Let's be honest. You wouldn't expect to learn to play baseball or chess by watching an expert do it. Math is the same — you'll need to do some work. BUT — with my calculator method, you'll see it's a whole new ball game!

THIS METHOD IS FRIENDLY! You can use it on your own. You follow interesting examples step-by-step with calculator keystrokes — instead of algebra.

Let me introduce myself — I'm Dr. George McCarty and I teach at the University of California. Tens of thousands of people like yourself have already used my guidebook to teach themselves calculus. Recently I have teamed up with another professor. We put those same tested and proven techniques to work for you in a new *Statistics Guidebook*.

Even if you've had trouble with math, this method will work for you. Choose the one you want to learn, calculus or statistics (or both) and start *today!*

As pennywise Ben Franklin said, "An investment in knowledge pays the best dividends."

CHOOSE STATISTICS

It's used everywhere — to set insurance rates and to assess a drug's effectiveness, to detect a crooked roulette wheel and to measure lightbulb lifetimes, to audit the books and to find your position at sea.

People have always known that their experience was not a perfect guide to the future. Statistics shows us that uncertainty is not only inevitable but it itself is measurable.

You'll find out about this in my guidebook. You'll read informal explanations with detailed calculations for topics like "confidence intervals" and "contingency tables."

YOU'LL BE DOING Inference and Hypothesis Testing, the Chi-Square Distribution, ANOVA, Regression and Correlation, Non-Parametric Methods — they're all here.

SUPPOSE YOU HAVE ALREADY STUDIED STATISTICS. You'll use Quick-Reference Guides to go straight to ■ topic you want to refresh yourself on — then get *right-now* answers in your own field, business, biology, science, psychology, education.

The Guidebook is guaranteed to work with any calculator or computer. If you would like to start out with a fancy new machine, order my Statistics Kit which includes the guidebook and a TI-35 calculator which finds means and standard deviations at the touch of a button. It has an LCD display, 1000 hour batteries — the works.

PHONE MC or VISA ORDERS TO
(714) 831-2637
Ask for our Catalog of books and calculators.

CHOOSE CALCULUS

The math of science and engineering — satellite orbits and compound interest, bacterial growth and radioactive decay, marginal profits and reaction rates.

You'll find out about these concepts when you use the Calculus Kit. This Kit contains my guidebook and a TI-35 calculator which finds log and trig values at the touch of a button. (It's modern, too, with LCD display and 1000 hour batteries.)

In the guidebook you'll learn both differential calculus and integral calculus — *even differential equations!*

NEED TO EVALUATE functions, areas, volumes — solve equations — use curves, trig, polar coordinates — find limits for sequences and series? It's all here!

If you're in the biological, social or physical sciences, you'll be doing Bessel functions, carbon dating, Gompertz' growth curves, half-life, future value, marginal costs, motion, cooling, probability, pressure — and plenty more.

Important numerical algorithms are here, too: rational and Padé approximation, bracketing, continued fractions, Euler's method, Heun's method, iteration functions, Newton's method, predictor-corrector, successive substitutions, Simpson's method and synthetic division.

If you already have a scientific calculator, you can get just the *Calculus Guidebook*.

George McCarty

P.S. — YOU *CAN* DO IT! I'm certain I can show you this easy way to success with math, so *I guarantee it to you unconditionally* — if you are not satisfied just send it back for a prompt refund!

LOOK AT WHAT THEY SAY...

- "...*excellent diagrams and clear explanations.*" —POPULAR SCIENCE
- Tektronix Engineer Bill Templeton says, *"The best, most clearly written book I have seen for improving your math skills."*
- "...*straight-to-the-point explanations...*" —HP Key Notes
- Professor H.I. Freedman of the U. of Alberta, states: *"There can be no question as to the usefulness of this book...lots of exercises...very clearly written...easy reading."*
- S.C. McCluney, Jr. of Philadelphia writes: *"With your book and ■ calculator the whole idea becomes clear in ■ moment, and is a MOST REFRESHING EXPERIENCE."*

TAX DEDUCTIBLE FOR PROFESSIONALS

YES, I want to learn math **fast** and **easy** with your *new method*. If I am not more than satisfied I can return it for a quick refund. RUSH me the guidebook or complete kit I select below (kit contains the guidebook and a modern TI-35 calculator).

☐ Send **FREE** Catalog of books & calculators

	STATISTICS GUIDEBOOK	KIT
How many?	☐ $14.95	☐ $39.95
Handling/Shipping	1.00	2.00
Tax (within Calif.)	.90	2.40
TOTAL		

	CALCULUS GUIDEBOOK	KIT
How many?	☐ $19.95	☐ $44.95
Handling/Shipping	1.00	2.00
Tax (within Calif.)	1.20	2.70
TOTAL		

***For AIR SHIPMENT** add $3.00 per item*

Phone (714)831-2637 (MC or VISA only)

Enclosed is ☐ Check, ☐ M.O.; or charge my MC/VISA... exp._____

Card #_____

Name _____

Address _____

City_____

State, Zip _____

OUTSIDE USA, **surface shipping is $3.00 per book or $5.00 per kit. Air shipment is $5.00 per book or $10.00 per kit. U.S. FUNDS PLEASE.**

EduCALC Publications, Dept. A5
27953 Cabot Road
South Laguna, CA 92677

©1982

THE AMATEUR SCIENTIST

What causes the color in plastic objects stressed between two polarizing filters?

by Jearl Walker

The failure of materials as a result of stress is both common and costly. One way to guard against such a failure is to test a sample object by applying enough stress to destroy it; when an identical object is put in service, the stresses are kept well below the destructive level. Another way is to make a model of the object out of plastic or some similar material and then to subject the model to photoelastic analysis, which yields a pictorial representation of the stresses in it.

Recently Frank R. Seufert of Cleveland showed me some examples of his photoelastic analysis of models of various objects. When someone is concerned about the stresses on an object, Seufert makes a model of it in Lexan or Tuffak plastic about an eighth of an inch thick. If the object is large, the model is a scaled-down version. Seufert mounts the model in a wood frame and undertakes to mimic the stress on the real object by stressing the model with rubber bands or machine screws. The screws work better because they can be tightened slowly and controlled more easily.

Seufert puts the model in front of a polarizing filter and illuminates it with light that passes from a 200-watt lamp through a ground-glass diffuser and then through the filter. Between the model and a 35-millimeter camera (a single-lens reflex model) he positions a long lens hood made out of cardboard. Attached to the camera is a Vivitar 2X teleconverter with its lenses removed. It serves as an extension tube to move the lens farther from the film; a different type of tube would serve as well. Seufert's setup includes a 135-mm. telephoto lens rated at *f*2.5. A second polarizing filter and a light blue filter (designated 80A) are mounted on the lens. The camera is held steady with the top piece from a tripod; the piece is attached to a block of plywood.

To make a photograph Seufert orients the first polarizing filter with its axis of polarization 45 degrees from the vertical. Then he puts his plastic model in front of it and turns on the lamp. As he sights through the camera he rotates the polarizing filter in front of the lens until a meaningful pattern is superposed on the image of the plastic model. He records the scene with a cable release that also triggers an electronic flash positioned near the lamp. The extra light is needed for a good photograph.

The film is Fuji color negative rated at ASA 100; Seufert has glossy prints made from the negatives. The colored lines in the accompanying photographs reveal the stress patterns in the plastic models. Regions under stress stand out. They are the places where the object is most likely to fail.

The first polarizing filter in Seufert's setup polarizes the light. The passage of this kind of light through stressed plastic encodes information about the stresses. The second polarizing filter (the one mounted on the camera) makes the coded information visible to an observer. The top illustration on page 149 depicts the stresses on a tiny, thin element that is part of a model being investigated by photoelasticity. Each edge of the element is under stress because of forces pulling perpendicular to the edge. Each edge is also under shear as material on opposite sides of the element attempt to slide in opposite directions.

Such is the nature of the stresses on a randomly chosen element. The picture is simplified if a differently oriented element is chosen at this region in the material. The new element is special in that it is square and has two important axes called the principal stress axes. The advantage of considering such a specially oriented element is that its edges are not affected by shear. The only stress on an edge is the perpendicular one. The orientation of the principal stress axes is what is revealed in a photograph of a stressed model illuminated with polarized light.

To understand the interaction of polarized light with the principal stresses in the plastic it is necessary to understand the nature of polarized light. According to classical physics, light is a wave composed of oscillating electric and magnetic fields. It is a peculiar kind of wave in that nothing material participates in the wave motion. Water waves accord more with intuition because a wave involves oscillations of the water surface. Something material participates. In the wave concept of light, however, the oscillation is of immaterial electric and magnetic fields.

The electric components specify the polarization of light. An electric field is a vector (of a specified size and direction) assigned to a point to be examined, as is shown in the middle illustration on page 149. The concept is useful when one considers how a charged particle might behave when it is placed at that point. It is also useful in forming a mental picture of light.

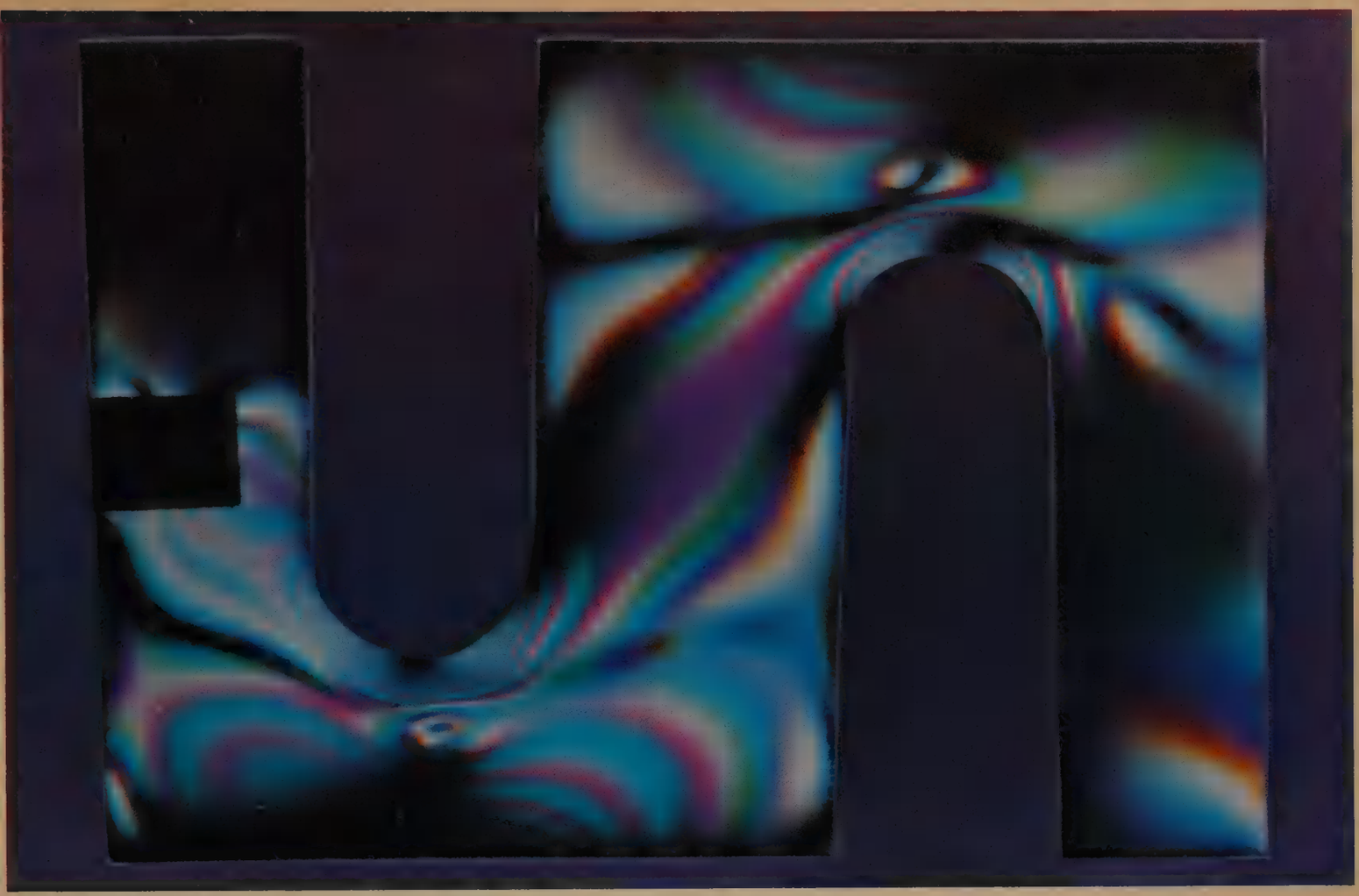
The illustration can be regarded as a snapshot of a light wave. Superposed on the ray, which indicates the direction of travel, are some of the electric vectors assigned to points along the ray. One point is singled out for scrutiny. At that point in the first snapshot the electric vector is long and points upward. A positive particle there would "feel" a strong force upward.

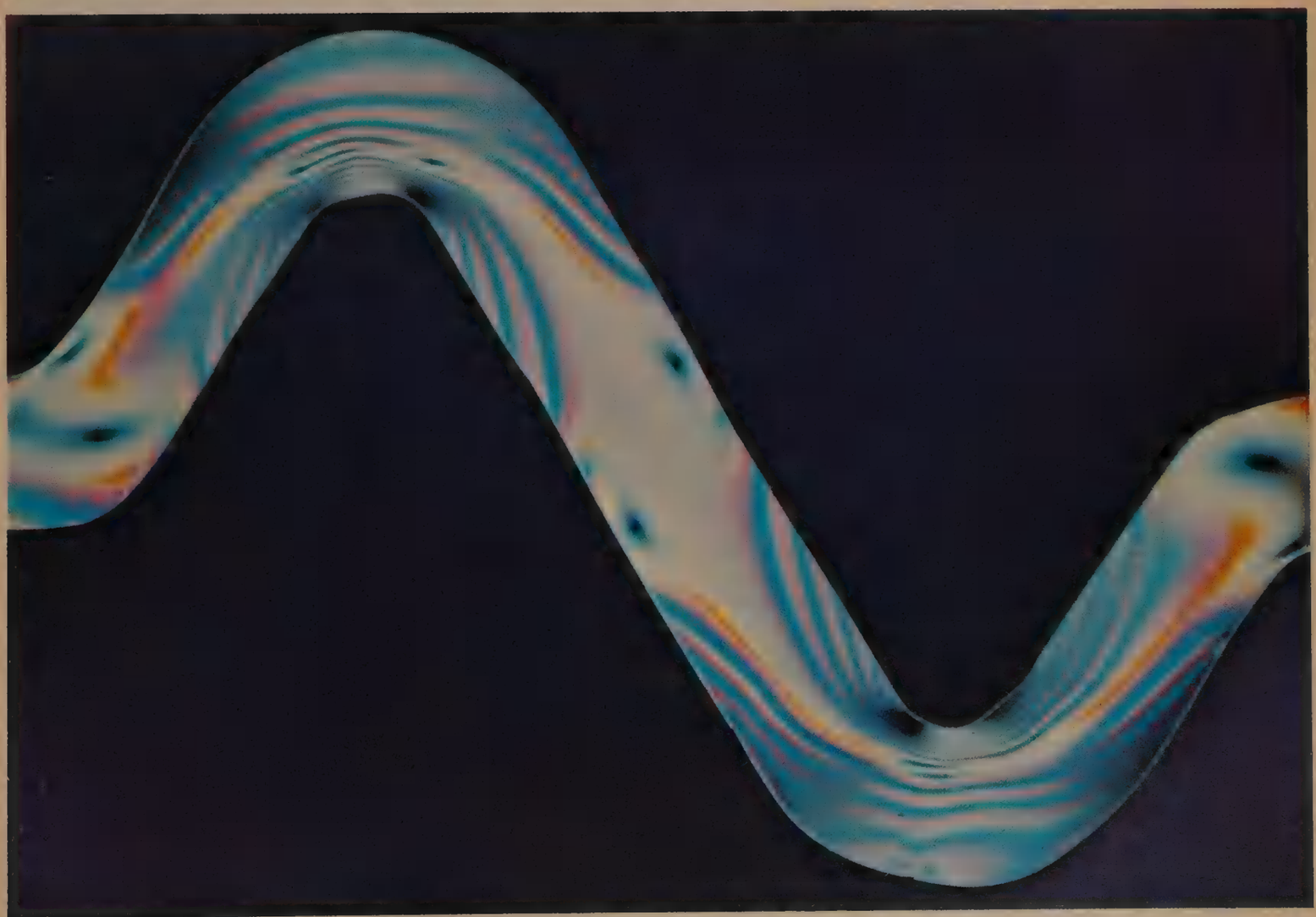
The light continues to move toward the right after the first snapshot is made. The electric field at the chosen point changes quickly, literally at the speed of light. A new snapshot reveals the change. The electric vector at the designated point is now downward. Since it is not one of the largest vectors in the illustration, the electric field is not as large as it could be and not as large as it will be immediately after this snapshot. A positive particle at this point would now "feel" a moderate downward force.

As light sweeps past the point, the electric vectors oscillate in direction and strength. Do not, however, be misled by the pictorial representation. The vectors are not material. A ray has no vectors sticking out of it like thorns on a rose stem. The electric vectors are bits of imagination stuck on a line called a ray.

Fictitious though the electric vectors are, they are almost essential to an understanding of the polarization of light. In the two snapshots of light the electric vectors are all flat in the plane of the page. Light from most common sources has vectors that are not as limited in direction. The vectors are necessarily perpendicular to a ray of the light, but they can point in any direction in a plane perpendicular to the ray. Such light is said to be unpolarized.

If light passes through a polarizing filter, the oscillation of the vectors is strictly limited to a single axis. Such light is said to be linearly polarized. (Some call it plane-polarized light.) If unpolarized light from an electric light bulb strikes a filter, the emerging light has its electric vectors confined to a single axis in a plane perpendicular to the ray. The direction of polarization of the light is the





Two of Frank R. Seufert's plastic models under compression

orientation of that axis. If the axis is vertical, the light is said to be vertically polarized.

Polarizing filters work by a process of elimination. The filter incorporates molecular chains that can be visualized as long, parallel absorbers. As light reaches them the electric vectors that oscillate parallel to them are eliminated; the perpendicular vectors pass through.

For example, if the filter's long molecules are stretched out horizontally (parallel to the x axis), only the horizontal components of the electric vectors in the light are eliminated. The bottom illustration on the opposite page presents a shorthand notation in which two double vectors represent unpolarized light; the filter eliminates the horizontal one and passes the vertical one. The result is vertically polarized light.

Normally the orientation of the molecules in a filter is not specified. Instead the filter is assigned a polarization axis perpendicular to the length of the molecules. This fictitious axis is parallel to the polarization of the emerging light.

Suppose a ray of vertically polarized light encounters a second polarizing filter. Will the light pass through? That depends on the polarization axis of the filter. If the axis is vertical, it is parallel to the polarization of the light approaching it. All the light is transmitted. If the axis is horizontal, none of the light is transmitted.

When light moving through air at a speed of 3×10^8 meters per second enters any transparent material, its effective speed decreases. The reason is that the light interacts with the molecules in its path. At each encounter with a molecule the light is absorbed; after a brief delay it is reemitted. Between molecules its speed is 3×10^8 meters per second, the value it has in a vacuum. Since it is intermittently delayed, however, it takes longer to travel through the material than it does to travel an equal distance in a vacuum. The light is said to be traveling slower in the material.

This effect was measured indirectly long before anyone knew about molecules. In order to tabulate the effect each transparent material was assigned the number called the index of refraction. A glass with an index of 1.6 transmits light slower than one with an index of 1.5. (In either case the actual time of transmission is so incredibly short that the difference is immaterial in everyday life.)

In 1816 David Brewster discovered how the index of refraction can aid in the analysis of stress in a transparent material. When he applied a stress to a sheet of glass illuminated with linearly polarized light, he found a change in the index of refraction of the glass. Moreover, the index then depended on how the light was polarized.

To follow Brewster's experiments visualize a vertical sheet of glass uniformly compressed by forces applied at the top and bottom in such a way that the principal stress axes are vertical and horizontal. When the light illuminating the glass is polarized vertically, it encounters a smaller index of refraction and moves with higher velocity than it does when it is polarized horizontally. If the glass is under tension (stretched by forces at the top and bottom), the results are just the opposite. A material of this type, with a speed of transmission that depends on the polarization of the light, is said to be birefringent, or doubly refracting.

A simple example will demonstrate how birefringence can aid in stress analysis. A vertical sheet of plastic is uniformly compressed to make the principal stress axes vertical and horizontal. The way polarized light interacts with the stressed plastic depends on whether the polarization of the light is aligned initially with one of those axes. You can create such an alignment by illuminating the plastic through a polarizing filter with its axis vertical. On the opposite side place another polarizing filter with its axis horizontal. The first filter is often called the polarizer, the second one the analyzer; the two are said to be crossed.

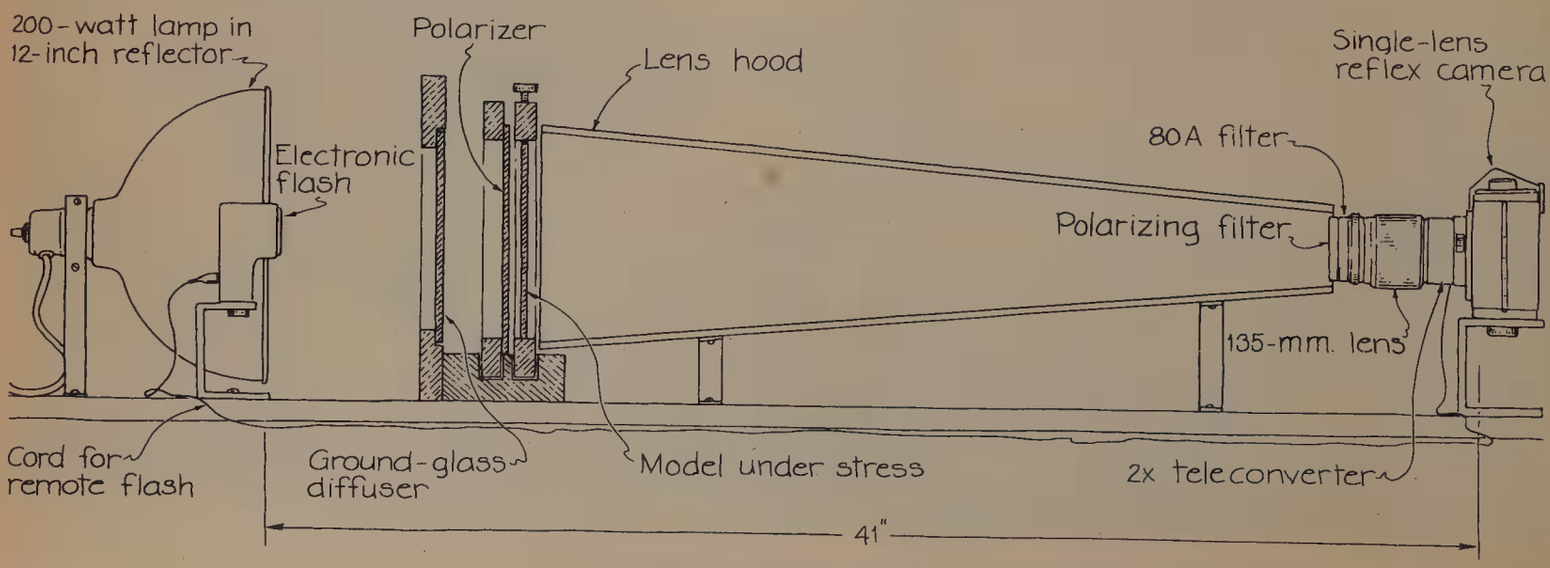
The light passes through the sheet of plastic at whatever speed the stress allows. The light emerging from the plastic is still vertically polarized and is therefore stopped by the analyzer. An observer looking through the analyzer sees darkness. In the experimental setup transmission is eliminated whenever the plastic is illuminated with light polarized parallel to a principal stress axis.

Next rotate each filter by 45 degrees in the same direction. The plastic sheet now receives light polarized at 45 degrees from the vertical. The course of light through the sheet is a bit harder to follow because the polarization must be considered in two components parallel to the principal stress axes. The two components travel through the plastic at different speeds because the indexes of refraction along the two axes differ.

When the components emerge from the plastic, they in effect recombine. The light is then likely to have a new direction of polarization. Whether or not the light will pass through the analyzer depends on how the polarization has been changed.

To determine the change one must consider how the wavelength of the light is altered by the plastic. For simplicity assume that the light is of a single wavelength. When light passes from air into a transparent material, its wavelength shortens. The larger the index of refraction, the shorter the wavelength of the light in the material.

Since stressed plastic has a different index for each principal stress axis, the extent to which the wavelength shortens depends on the polarization of the light. If the light is vertically polarized, its wavelength is divided by the index of refraction associated with the vertical principal stress axis. If the light is horizontally polarized, the wavelength is divided by the index associated with the horizontal axis. When the polarization lies between the two axes, it must be considered in components. The vertical



Seufert's experimental setup

component is shortened in wavelength by one amount and the horizontal component by a different amount.

As the two components proceed through the plastic with their different wavelengths they oscillate a different number of times. For example, the component polarized parallel to the vertical stress axis might have 1,000 oscillations (wavelengths); the other component, being of shorter wavelength, might have an additional oscillation, for a total of 1,001. The two components begin their trip exactly in step and emerge again in step even though one of them has an additional oscillation.

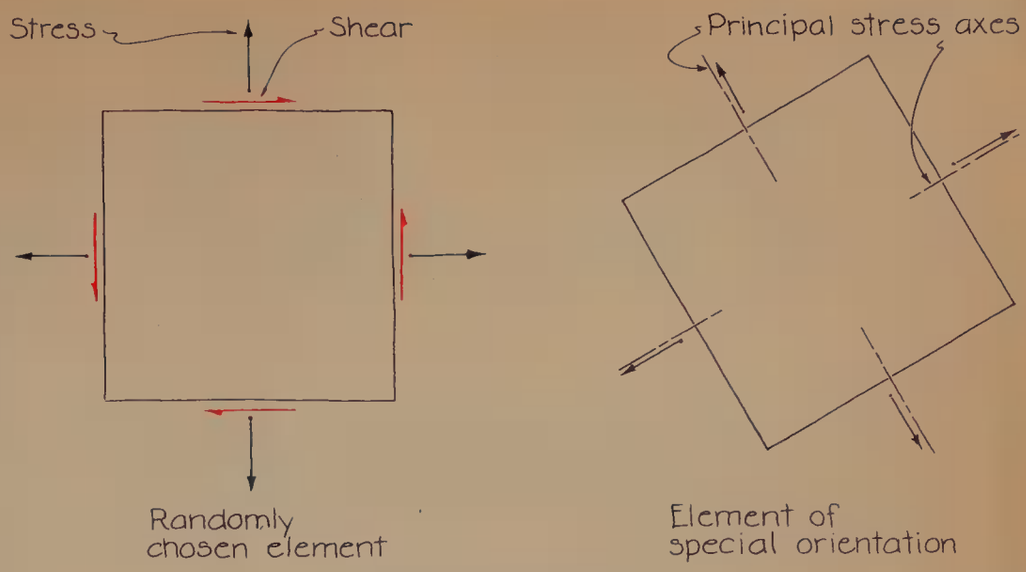
When one mathematically recombines the two emerging components, the electric vectors of the light are found to be oscillating exactly as they were before they entered the plastic. The polarization is tilted from the vertical by 45 degrees. Since the polarizing filters are crossed, the light reaching the analyzer is blocked.

If instead one component has an additional half oscillation, recombination yields a polarization rotated by 90 degrees. This new polarization is parallel to the polarization axis of the analyzer and so the light passes through.

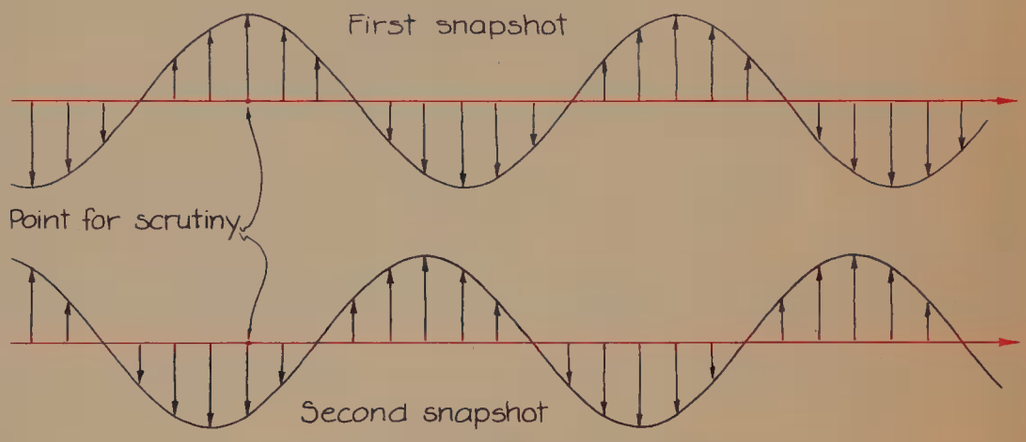
Any intermediate result is also possible. Recombination is then harder to visualize. The components do not yield linearly polarized light. Instead the polarization rotates continuously about the ray of light. The electric vector rotates from upward to downward and then back to upward. Light with a rotating polarization is said to be elliptically polarized. (In the special case where the maximum size of the electric vector remains the same throughout the rotation, the light is said to be circularly polarized.) Elliptically polarized light is partly transmitted by the analyzer. The component of polarization that is parallel to the polarization axis of the analyzer is passed whereas the other component is blocked.

The point of this discussion is that what the observer sees through the analyzer depends on the angle between the polarization of the light and a principal stress axis in the plastic. Alignment of the polarization with an axis yields darkness at the analyzer. If in misalignment the plastic releases light of unchanged polarization, the analyzer again blocks it. In other instances at least some light gets through the analyzer.

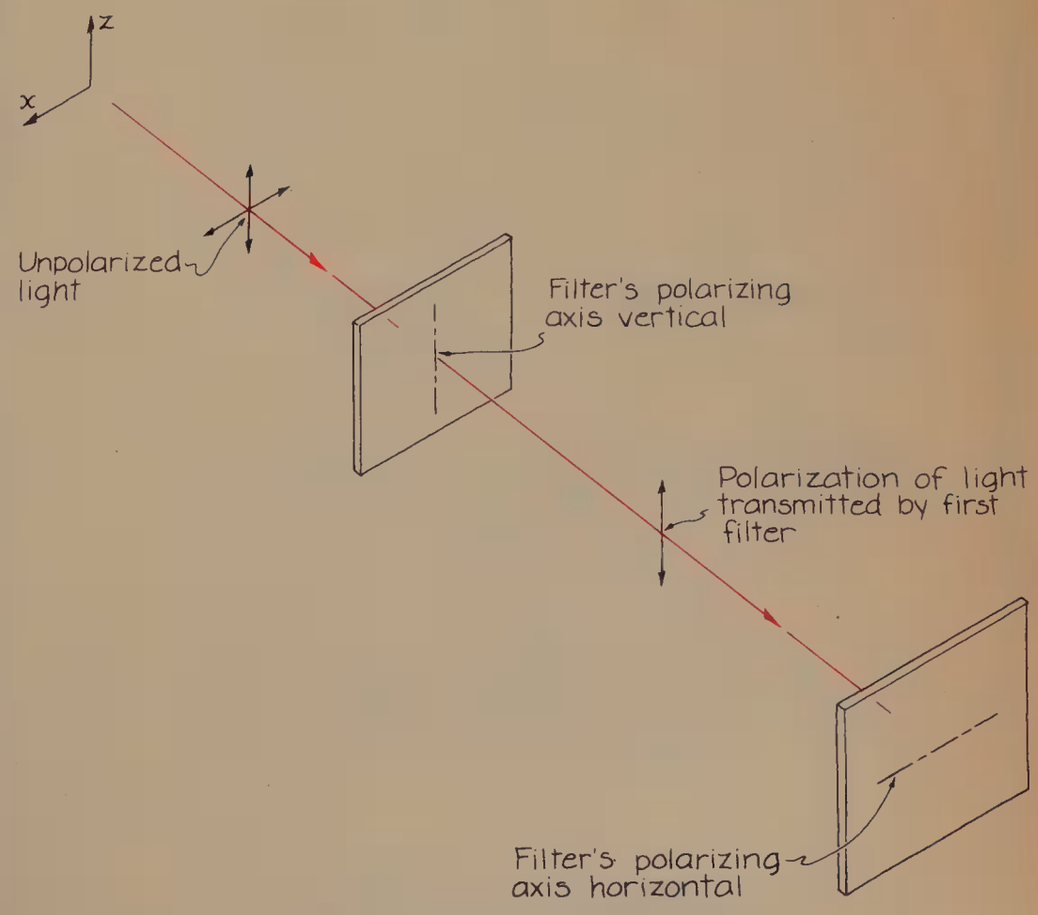
The example is simple in that the stress is uniform and the principal stress axes are vertical and horizontal at any point in the plastic sheet. If the plastic is subjected to various stresses, the axes will be oriented differently for each point in it. The purpose of photoelasticity is to discover how they are oriented; in this way one might be able to locate a section in the model that might fail under the stress. When a complex and



Stress and shear on an element



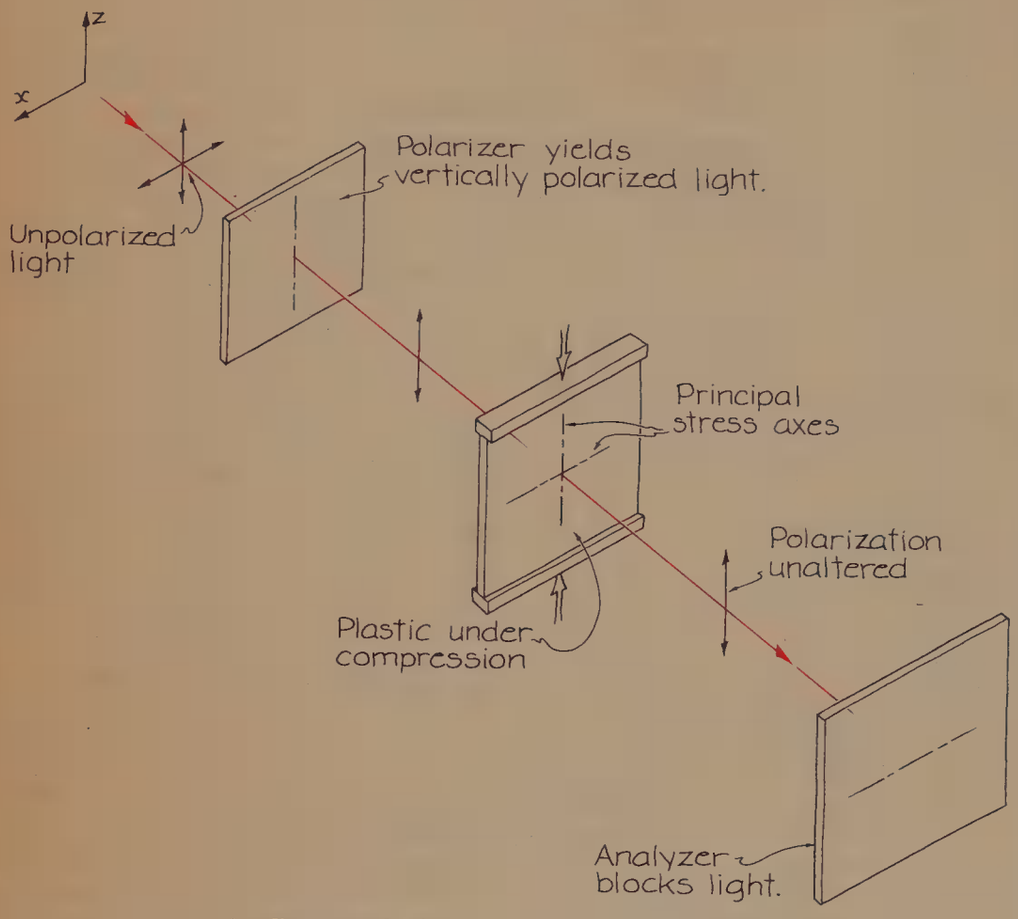
"Snapshots" of electric vectors along a ray of light



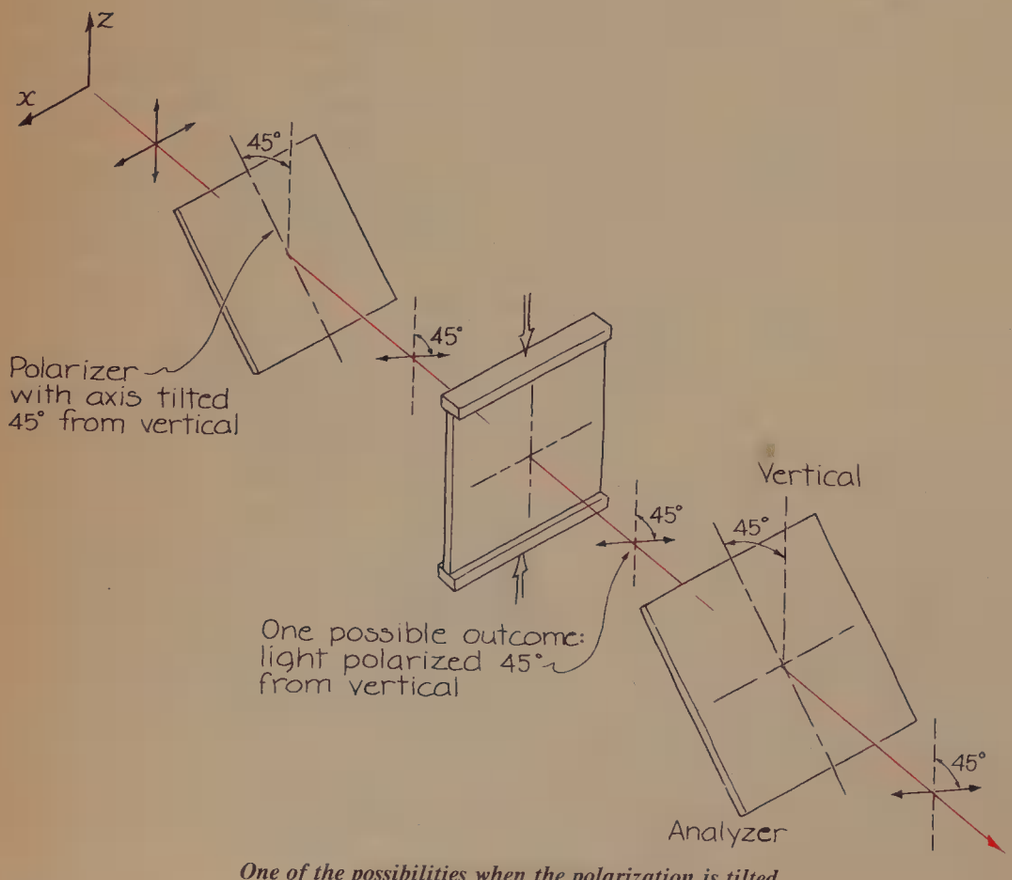
How crossed polarizing filters block light

unknown arrangement of stresses is applied to the plastic, the model as seen through the analyzer has a superposed pattern of the dark and bright lines called fringes.

A dark fringe marks points within the plastic from which the emerging light is polarized in exactly the way that keeps it from going through the analyzer. The reason is either that the light passing through one of these dark points is polarized parallel to one of the principal axes or that it has two components (one parallel to each principal stress axis there) that recombine to yield the same polarization the light had when it entered the plastic. In either case the analyzer blocks the light.



An alignment of polarization with a principal stress axis



One of the possibilities when the polarization is tilted

Usually the first condition creates most of the dark fringes. A fringe caused in this way is called an isoclinic. The pattern of isoclinics reveals the orientation of the principal stress axes within the plastic.

To map the principal axes you photograph the plastic through the analyzer at a particular orientation of the two polarizing filters. Suppose the axis of the first filter is vertical and the axis of the analyzer is horizontal. The plastic is thus illuminated with vertically polarized light. The isoclinics photographed through the analyzer mark the points on the plastic that have one of their principal stress axes vertical. (Since the stress axes are at right angles to each other, the other axis must then be horizontal.) The pattern in the photograph is traced on paper, and the principal axes are drawn superposed on several points on the isoclinics.

Next the filters are rotated 10 degrees, say, and another photograph is made. The isoclinics mark the points with a principal stress axis tilted from the vertical by 10 degrees. These isoclinics are added to the trace and again the principal axes are sketched in at a few places. After a few more photographs have been made the trace shows how the principal stress axes are oriented at many places in the plastic.

Now you add to the map lines that connect points having equal stress values. For example, a line would be drawn from a principal axis at one point to another principal axis at a nearby point. It is likely to be a curved line. Although a certain amount of guesswork is inescapable here, the result is a rough map of the lines of principal stress.

The dark fringes resulting from recombination are called isochromatics. They are usually masked by the isoclinics, but they can be employed to assign values to the principal stresses revealed by the isoclinics. To make use of them you must first eliminate the isoclinics.

You can do it by putting two additional filters, called quarter-wave plates, in the path of the light. One plate goes between the first polarizer and the plastic and the other one goes between the plastic and the analyzer. The function of a quarter-wave plate is to transform linearly polarized light into circularly polarized light.

The plate works something like the stressed plastic being analyzed. The plate is birefringent, that is, it has two orthogonal axes ("fast" and "slow") that pass light at different speeds. Suppose the first polarizer yields light that is polarized vertically. The first quarter-wave plate is set with its fast axis tilted from the vertical by 45 degrees. The light reaching the plate thus has components along both the fast and the slow axes. As the two components pass through the plate they oscillate a different number

of times. The polarization of the light emerging from the plate depends on that difference. Since the plates are designed to make one component complete a quarter of an oscillation more than the other component (hence the name quarter-wave plate), circularly polarized light emerges.

The second quarter-wave plate is arranged with its fast axis perpendicular to that of the first one. The function of the second plate is to subtract the quarter-wavelength difference imparted to the two components of light by the first plate. The net result may seem hardly worth the trouble, but a useful end is gained. When the circularly polarized light passes through the stressed plastic, the polarization cannot be strictly parallel to a principal stress axis. Therefore the isoclinics resulting from such an alignment are eliminated. What one receives through the analyzer is a pattern of the isochromatics.

The advantage of this pattern is that the fringes are related to the size of the stresses in the plastic. At any given point in the plastic the difference between the principal stresses determines the polarization of the light. Thus the stress difference determines whether the point ends up as part of a dark fringe or as part of a bright one. In principle the stress values for any point in the plastic can be computed by examining the isochromatics.

Since the isochromatic pattern depends on the shortening of the wavelengths of the two components of light transmitted through any point in the plastic, the pattern depends on the wavelength of the light illuminating the plastic. With white light each color creates its own pattern. At one point in the plastic one of the colors might leave with the same polarization it had when it entered. That color is eliminated by the analyzer. The other colors passing through the same point are changed somewhat in polarization and so are at least partly transmitted by the analyzer. The observer sees that point on the plastic as being neither white nor dark but a color. The composite of the points generating the same color forms a pattern of colored isochromatics.

The actual color the observer will see at a particular point is difficult to predict. It depends on how the various colors are altered in polarization. It also depends on the perception of the observer or the color response of the film in a camera. Color prediction becomes even more complicated when the stress differences at each point in the plastic increase. When the differences are large enough, the colors begin to wash out into white.

The photographs made by Seufert are from experiments where linearly polarized light is sent through stressed plastic. Both the isoclinics (which are only bright and dark) and the isochromatics

KING FAISAL INTERNATIONAL PRIZE

General Secretariat

AN INVITATION TO THE NOMINATION FOR THE KING FAISAL INTERNATIONAL PRIZE IN MEDICINE

The General Secretariat of The King Faisal International Prize, at Riyadh, Kingdom of Saudi Arabia, has the honour to invite the Universities, Academies, Educational Institutions and Research Centers all over the World to nominate qualified candidates for The King Faisal International Prize in Medicine to be awarded in Rabi-al-Awal 1404 AH. i.e. January 1984.

(a) The topic of the prize will be "DIARRHEAL DISEASES".

(b) Selection shall be according to the discretion and decision of a Committee consisting of National and International assessors selected by The Board of King Faisal International Prize.

(c) More than one person may share the prize.

(d) The winners' names will be announced in January 1984, and the prize will be awarded in an official ceremony to be held for that purpose in Riyadh, Kingdom of Saudi Arabia.

(e) The Prize consists of:

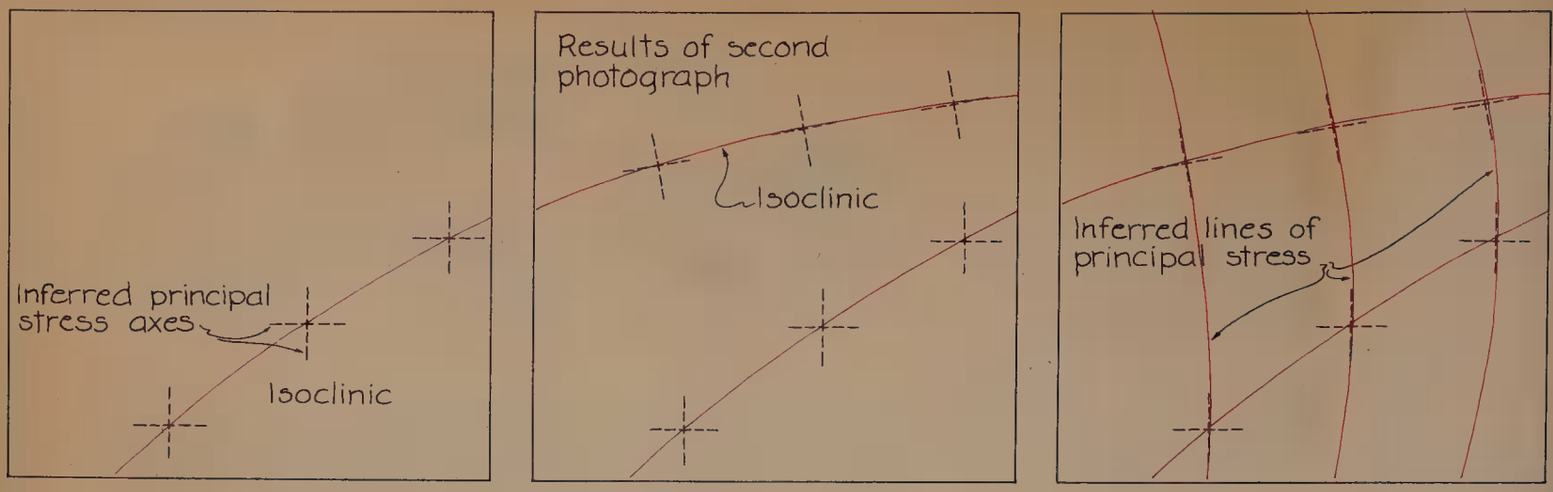
(1) A certificate in the name of the winner containing abstract of his work that qualified him for the prize.

(2) A precious medal.

(3) A sum of two hundred fifty thousand Saudi Riyals (SR. 250,000).

(f) Nominees should satisfy the following conditions :

1. A nominee must have accomplished an outstanding academic work in the subject of the prize leading to the benefit of mankind and enrichment of human thought.
2. The work submitted with the nomination of the prize must have already been printed and published. If possible an abstract in Arabic should be attached if the work was published in any other language.
3. The prize will be awarded for specific original research but the life-time background of the work will be taken into account.
4. The specific work submitted must not have been awarded a prize by any international educational institution, scientific organization, or foundation.
5. The nomination must be submitted by leading members of recognised educational institutions and of world-fame, such as Universities, Academies and Research Centers. The nominations of other individuals and political parties will not be accepted.
6. The nominations must give full particulars of the nominee's academic background, experiences and/or his publications, copies of his educational certificates, if available, and three 6 x 9 cm photographs. The nominee's full address and telephone number are also requested.
7. The nominations and works in ten copies are to be sent by registered air mail to the address stated below.
8. The latest date for receipt of the full nominations with copies of works is the 12th of Dhu Al-Qe'dah, 1403 AH. i.e. the 20th of August 1983. The nomination papers received after this date will not be considered unless the subject of the prize is postponed to the following year.
9. No nomination papers or works will be returned to the senders.
10. Enquiries should be made, and nominations should be sent, to the Secretary General of the King Faisal International Prize, P. O. Box 22476, Riyadh 11495, Kingdom of Saudi Arabia, Telex 204667 PRIZE SJ.

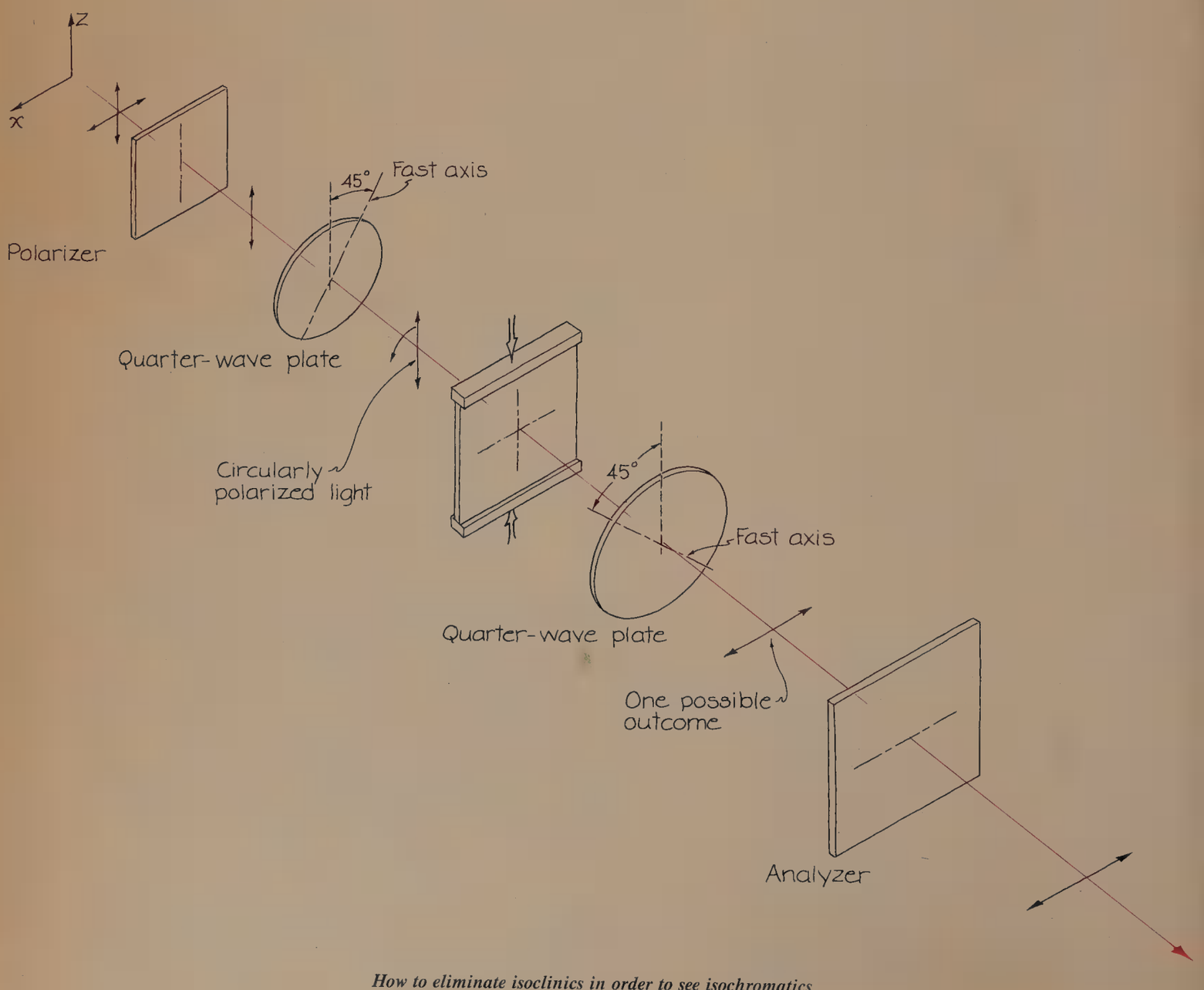


The analysis of isoclinics

(which are color-dependent) are included in the photographs.

The two photographs on page 147 show plastic models with concentrations of stress in their concave regions. The isochromatics are closely spaced there, indicating that the stress differences at points in those regions vary considerably. Since the models are under compression from forces on both the left and the right, the concave regions are under compression. In the top photograph the convex region is being stretched. The more uniform distribution of color in the straight sections indicates that the distribution of stress is more uniform there than it is in the curved sections.

The polarizing filters and quarter-wave plates for photographs of this type can be obtained from the Edmund Scientific Co. (101 East Gloucester Pike, Barrington, N.J. 08007) or Jerryco, Inc. (601 Linden Place, Evanston, Ill. 60202). Readers wanting to write to Seufert may do so at 2050 West Boulevard, Cleveland, Ohio 44102.



How to eliminate isoclinics in order to see isochromatics

THE INPUT/OUTPUT STRUCTURE OF THE UNITED STATES ECONOMY

WHAT MAKES THE U.S. ECONOMY TICK?

The editors of SCIENTIFIC AMERICAN have prepared a wall chart displaying for the 1980's the Input/Output Structure of the U.S. Economy based on the latest interindustry study from the U.S. Department of Commerce.

The SCIENTIFIC AMERICAN Input/Output wall chart does for economics what the table of elements does for chemistry. It answers at a glance questions about the linkage between the microeconomics of the firm and the macroeconomics of the system; about the web of technological interdependencies that tie industry to industry; about the industry-by-industry direct and indirect consequences of swings in public and private spending; about the impact of change in technology, and about any other topic you can think of. You are rewarded by surprise as well as by confirmation of your hunches. For teaching and practical and theoretical studies, here is a powerful, graphic tool.

In the familiar format of the SCIENTIFIC AMERICAN Input/Output wall charts for the 1960's and 1970's, the wall chart for the 1980's measures 65″ × 52″ and is printed in eight colors. Each of the nearly 10,000 cells in the 97-sector interindustry matrix shows (1) the interindustry commodity flow, (2) the direct input/output coefficient and (3) the "inverse" coefficient. Where the direct input/output coefficient exceeds .01, the cell is tinted in the color code of the industrial bloc from which the input comes. This device, combined with triangulation of the matrix, brings the structure of interindustry transactions into graphic visibility.

A supplementary table displays, industry by industry, the capital stock employed; the employment of managerial, technical-professional, white-collar and blue-collar personnel; the energy consumption by major categories of fuel, and environmental stress measured by tons of pollutants.

The editors of SCIENTIFIC AMERICAN are happy to acknowledge the collaboration, in the preparation of this wall chart, of Wassily Leontief, originator of input/output analysis—for which contribution to the intellectual apparatus of economics he received the 1973 Nobel prize—and director of the Institute for Economic Analysis at New York University.

Packaged with the chart is an index showing the BEA and SIC code industries aggregated in each of the 97 sectors.

SCIENTIFIC AMERICAN

415 Madison Avenue, Dept. I/O, New York, N.Y. 10017

YES, send me the new SCIENTIFIC AMERICAN 1980's INPUT/OUTPUT wall chart—printed in eight colors on heavy paper measuring 65″ × 52″.

I understand my chart(s) will be shipped prepaid, uncreased*, flat-rolled and in a sturdy mailing tube. (Allow 4-6 weeks for delivery)

I enclose a check for $____ for ____ chart(s) at $35 each, saving me $5 each in handling charges.

Bill me for $____ for ____ chart(s) at $40 each.

Name_____

Company_____

Street Address_____

City, State, Zip Code_____

*Charts shipped international require one fold. 6/83

Make check payable to SCIENTIFIC AMERICAN. Please add sales tax for California, Illinois, Massachusetts, Michigan and New York.

BIBLIOGRAPHY

Readers interested in further explanation of the subjects covered by the articles in this issue may find the following lists of publications helpful.

METAMAGICAL THEMAS

ESCAPE FROM PARADOX. Anatol Rapoport in *Scientific American,* Vol. 217, No. 1, pages 50–56; July, 1967.

NEWCOMB'S PROBLEM AND THE PRISONER'S DILEMMA. Steven J. Brams in *The Journal of Conflict Resolution,* Vol. 19, No. 4, pages 596–612; December, 1975.

THE PRISONERS OF INSECURITY. Bruce Russett. W. H. Freeman and Co., 1983.

PHYSICAL DISABILITY AND PUBLIC POLICY

DESIGN FOR INDEPENDENT LIVING: THE ENVIRONMENT AND PHYSICALLY DISABLED PEOPLE. Raymond Lifchez and Barbara Winslow. University of California Press, 1979.

ENVIRONMENTAL ACCESSIBILITY AND INDEPENDENT LIVING OUTCOMES: DIRECTIONS FOR DISABILITY POLICY AND RESEARCH. Gerben DeJong. University Center for International Rehabilitation, Michigan State University, 1981.

DISABLED PEOPLE AS SECOND CLASS CITIZENS. Edited by Myron Eisenberg. Cynthia Griggins and Richard Duval. Springer Publishing Company, 1982.

A VECTOR FOR INTRODUCING NEW GENES INTO PLANTS

PLASMIDS SPECIFYING PLANT HYPERPLASIAS. Eugene W. Nester and Tsune Kosuge in *Annual Review of Microbiology,* Vol. 35, pages 531–565; 1981.

MOLECULAR BIOLOGY OF PLANT TUMORS. Edited by G. Kahl and J. S. Schell. Academic Press, 1982.

T-DNA OF THE AGROBACTERIUM TI AND RI PLASMIDS. Michael Webster Bevan and Mary-Dell Chilton in *Annual Review of Genetics,* Vol. 16, pages 357–384; 1982.

PROSPECTS IN PLANT GENETIC ENGINEERING. Kenneth A. Barton and Winston J. Brill in *Science,* Vol. 219, No. 4585; pages 671–676; February 11, 1983.

GIANT VOLCANIC CALDERAS

RESURGENT CAULDRONS. Robert L. Smith and Roy A. Bailey in *The Geological Society of America: Memoir 116,* pages 613–662; 1968.

THEORETICAL MODELING OF THE GENERATION, MOVEMENT, AND EMPLACEMENT OF PYROCLASTIC FLOWS BY COLUMN COLLAPSE. R. S. J. Sparks, L. Wilson and G. Hulme in *Journal of Geophysical Research,* Vol. 83, No. B4, pages 1727–1739; April 10, 1978.

YELLOWSTONE PARK AS A WINDOW ON THE EARTH'S INTERIOR. Robert B. Smith and Robert L. Christiansen in *Scientific American,* Vol. 242, No. 2, pages 104–117; February, 1980.

PEKING MAN

ON A LOWER MOLAR HOMINID TOOTH FROM THE CHOUKOUTIEN DEPOSIT. Davidson Black in *Palaeontologia Sinica,* Series D, Vol. 7, Fascicule 1, pages 1–28; 1927.

ON AN ADOLESCENT SKULL OF *SINANTHROPUS PEKINENSIS* IN COMPARISON WITH AN ADULT SKULL OF THE SAME SPECIES AND WITH OTHER HOMINID SKULLS, RECENT AND FOSSIL. Davidson Black in *Palaeontologia Sinica,* Series D, Vol. 7, Fascicule 2, pages 1–144; 1927.

NOTICE OF THE DISCOVERY OF QUARTZ AND OTHER STONE ARTIFACTS IN THE LOWER PLEISTOCENE HOMINID-BEARING SEDIMENTS OF THE CHOUKOUTIEN CAVE DEPOSIT. Pei W. C. in *Bulletin of the Geological Society of China,* Vol. 11, No. 2, pages 109–139; December, 1931.

THE *SINANTHROPUS* POPULATION OF CHOU KOU TIEN. Franz Weidenreich in *Bulletin of the Geological Society of China,* Vol. 14, No. 4, pages 427–461; 1935.

THE CAVE HOME OF PEKING MAN. Chia Lan-po. Foreign Language Press, 1975.

DARK MATTER IN SPIRAL GALAXIES

DYNAMIC EVIDENCE ON MASSIVE CORONAS OF GALAXIES. Jaan Einasto, Ants Kaasik and Enn Saar in *Nature,* Vol. 250, No. 5464, pages 309–310; July 26, 1974.

MASSES AND MASS-TO-LIGHT RATIOS OF GALAXIES. S. M. Faber and J. S. Gallagher in *Annual Review of Astronomy and Astrophysics,* Vol. 17, pages 135–187; 1979.

THE LARGE-SCALE STRUCTURE OF THE UNIVERSE. J. P. E. Peebles. Princeton University Press, 1980.

THE PHYSIOLOGICAL ECOLOGY OF WHALES AND PORPOISES

PHYSIOLOGY OF A SMALL CETACEAN. J. Kanwisher and G. Sundnes in *Hvalradets Skrifter,* No. 48, pages 45–53; 1965.

THERMAL REGULATION IN CETACEANS. John Kanwisher and Gunnar Sundnes in *Whales, Dolphins and Porpoises,* edited by Kenneth S. Norris. University of California Press, 1966.

RESPIRATION AND DEEP DIVING IN THE BOTTLENOSE PORPOISE. Sam H. Ridgway, B. L. Scronce and John Kanwisher in *Science,* Vol. 166, No. 3913, pages 1651–1654; December 26, 1969.

DOLPHIN LUNG COLLAPSE AND INTRAMUSCULAR CIRCULATION DURING FREE DIVING: EVIDENCE FROM NITROGEN WASHOUT. Sam H. Ridgway and Red Howard in *Science,* Vol. 206, No. 4423, pages 1182–1183; December 7, 1979.

FREE AND FORCED DIVING IN BIRDS. John W. Kanwisher, Geir Gabrielsen and Nancy Kanwisher in *Science,* Vol. 211, No. 4483, pages 717–719; February 13, 1981.

GEORG CANTOR AND THE ORIGINS OF TRANSFINITE SET THEORY

WHAT IS CANTOR'S CONTINUUM PROBLEM? Kurt Gödel in *Philosophy of Mathematics: Selected Readings,* edited by Paul Benacerraf and Hilary Putnam. Prentice-Hall, Inc., 1964.

PROBLEME DES UNENDLICHEN: WERK UND LEBEN GEORG CANTORS. Herbert Meschkowski. Vieweg, Braunschweig, 1967.

TOWARDS A BIOGRAPHY OF GEORG CANTOR. I. Grattan-Guinness in *Annals of Science,* Vol. 27, No. 4, pages 345–391; December, 1971.

GEORG CANTOR: HIS MATHEMATICS AND PHILOSOPHY OF THE INFINITE. Joseph Warren Dauben. Harvard University Press, 1979.

THE SLIDE FASTENER

THE INVENTION AND INTRODUCTION OF THE ZIPPER. P. J. Federico in *Journal of the Patent Office Society,* Vol. 28, No. 12, pages 855–876; December, 1946.

THE SLIDE FASTENER INDUSTRY IN 1950: A REPORT FOR SLIDE FASTENER ASSOCIATION, INC. Prepared by Business Information Division, Marketing Services Company. Dun and Bradstreet, Inc., 1951.

TALON, INC.: A ROMANCE OF ACHIEVEMENT. James Gray. Meadville Printcraft Corp., 1963.

THE AMATEUR SCIENTIST

PHOTOELASTICITY: PRINCIPLES & METHODS. H. T. Jessop and F. C. Harris. Dover Publications, Inc., 1960.

POLARIZED LIGHT. William A. Shurcliff and Stanley S. Ballard. D. Van Nostrand Company, Inc., 1964.

PHOTO-ELASTIC ANALYSIS. A. W. Hendry. Pergamon Press, 1966.

MARYGROVE
LIBRARY
Detroit, MI.